Technical co-Sponsor :

ISBN :
978-1-5386-7485-7
IEEE CATALOG NUMBER :
CFP18CUE-USB

2018 iSemantic

3rd International Seminar on Application for Technology of Information and Communication

PROCEEDINGS

Creative Technology for Human Life

Semarang | September 21st - 22nd, 2018

Email:
isemantic@lppm.dinus.ac.id
Website:
http://isemantic.dinus.ac.id/

Technically Co-sponsored by:

IEEE

isemantic

International Seminar on Application for Technology of Information and Communication

UNIVERSITAS DIAN NUSWANTORO

UDINUS

SEMARANG

PROCEEDINGS

Creative Technology for Human Life



ISEMANTIC 2018

3rd International Seminar on Application for Technology of Information and Communication

September 21st – 22nd, 2018

Universitas Dian Nuswantoro

Semarang, Indonesia

ISBN : 978-1-5386-7485-7
IEEE CATALOG NUMBER : CFP18CUE-USB

2018 International Seminar on Application for Technology of Information and Communication

(iSemantic)

PROCEEDINGS

Creative Technology for Human Life

September 21st – 22nd, 2018

Universitas Dian Nuswantoro

Semarang, Indonesia

ISBN : 978-1-5386-7485-7
IEEE CATALOG NUMBER : CFP18CUE-USB

COPYRIGHT

2018 International Seminar on Application for Technology of Information and Communication (iSemantic)

Copyright© 2018 by the Institute of Electrical and Electronics Engineers, Inc. All right Reserved

Copyright and Reprint Permission:
Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law, for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

Other copying, reprint, or reproduction requests should be addresses to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, P.O. Box 1331, Piscataway, NJ 08855-1331.

IEEE Catalog Number: CFP18CUE-USB
ISBN: 978-1-5386-7485-7

TABLE OF CONTENTS

TABLE OF CONTENTS iv
GREETINGS FROM THE GENERAL CHAIR v
GREETINGS FROM RECTOR vi
COMMITTEE vii
GENERAL INFORMATION ix
VENUE ix
MAPS ix
KEYNOTE SPEAKER 1 x
KEYNOTE SPEAKER 2 xi
KEYNOTE SPEAKER 3 xii
PROCEEDINGS xiii

GREETINGS FROM THE GENERAL CHAIR

Dearest Conference Participants!

We have gathered here to discuss the scope of Application for Technology of Information and Communication. I am highly honored to take this opportunity, on behalf of all committee and organizer, to welcome all of you today to this 3rd International Seminar on Application for Technology of Information and Communication (iSemantic 2018)

We have two previous seminars which are iSemantic 2016 and iSemantic 2017, those were a great success. All the papers presented last year have been successfully published in IEEE Xplore and indexed by SCOPUS. And now, in our third with new topic “Creative Technology for Human Life”, there are 112 papers accepted out of 170 submitted papers and are ready to be presented by the authors in the conference today.

It is also my honor today to have three great persons who have agreed to be the Keynote Speakers for this seminar. The distinguished 3 persons are as follows:

- Prof. Dr. Nurul Taufiqu Rochman from Indonesian Institute of Sciences, Indonesia
- Prof. Peter Wushou Chang, MD, MPH, ScD from Tufts University School of Medicine, Taiwan
- Reece Hinchcliff, Ph.D. from University of Technology Sydney, Australia

Therefore, I would like to thank all the organizers, supporters, and organizing committee members of iSemantic 2018 who have made this conference happen. I believe that without their support and hard work, iSemantic 2018 would not have come true.

We do hope that this conference is also going to be a great success like the two last years, and we request for your support to make this event a memorable and successful one.

With warm regards,



Dr. Guruh Fajar Shidik, S.Kom., M.Cs

General Chair iSemantic 2018

Universitas Dian Nuswantoro

GREETINGS FROM RECTOR

Dearest Conference Participants!

Welcome to Semarang, Indonesia. Welcome to the 3rd International Seminar on Application for Technology of Information and Communication (iSemantic 2018). As the Rector of Universitas Dian Nuswantoro and on behalf of the Organizing Committee of iSemantic 2018, I would like to say that I am honored to have all of you here in this wonderful event.

In this occasion, I would like to thank you all and welcome Prof. Dr. Nurul Taufiqu Rochman from Indonesian Institute of Sciences, Indonesia, Prof. Peter Wushou Chang, MD, MPH, ScD from Tufts University School of Medicine, Taiwan, and also Mr. Reece Hinchcliff, Ph.D. from University of Technology Sydney, Australia, for your being here and for being Keynote speakers for today's conference.

And I would like to warmly welcome all the participants and the authors of this 3rd International Seminar on Application for Technology of Information and Communication. With the new topic "Creative Technology for Human Life". I believe that through this great conference and the ideas from all the authors, the quality of human life could be improved.

And for that reason, I would like to encourage all of you as academia, researchers and industrial society to keep improving the quality of human life and the world by empowering technology and science.

I am sure that you all will feel enriched with knowledge after attending this conference. I welcome you all once again to this event and hope that you all will have a great time ahead.

With warm regards,



Prof. Dr. Ir. Edi Noersasongko, M.Kom.

Rector of Universitas Dian Nuswantoro

COMMITTEE

General Chair :

Dr. Pulung Nurtantio Andono, M.Kom (Universitas Dian Nuswantoro - Indonesia)

Steering Committee :

Prof. Dr. Ir. Edi Noersasongko, M.Kom (Universitas Dian Nuswantoro - Indonesia)
Prof. Supriyadi Rustad (Universitas Dian Nuswantoro - Indonesia)
Prof. Dr. Taufik (Calpoly – USA)
Prof. Kunio Kondo (Tokyo University of Technology – Japan)
Prof. Kiyoaki Aikawa (Tokyo University of Technology - Japan)
Prof. Taichi Watanabe (Tokyo University of Technology Japan)
Prof. Masanori Kakimoto (Tokyo University of Technology - Japan)
Prof. Vincent Didik Wiet Aryanto (UDINUS - Indonesia)
Prof. Sumonta Kasemvillas (Khon Kaen Univ - Thailand)
Prof. Ahmad Yusoff (USM - Malaysia)
Prof. Mohamad Ashari (Institut Teknologi Sepuluh November - Indonesia)
Prof. Nana Suryana Herman (UTeM - Malaysia)
Prof. Mauridhi Hery Purnomo (Institut Teknologi Sepuluh November - Indonesia)
Prof. Benyamin Kusumoputro (Universitas Indonesia - Indonesia)
Prof. Alen Soldo (University of Split- Croatia)
Prof. Zaenal A. Hasibuan (Universitas Indonesia - Indonesia)
Prof. Bambang Riyanto (Institut Teknologi Bandung - Indonesia)
Prof. Ida Ayu Giriantari (Universitas Udayana – Indonesia)
Prof. Andreas Lako (Unika Soegijopranoto – Indonesia)
Prof. Adang Suwandi (Institut Teknologi Bandung - Indonesia)
Prof. Yanuarsyah Haroen (Institut Teknologi Bandung - Indonesia)
Prof. Sasongko Pramono Hadi (Universitas Gajah Mada - Indonesia)
Dr. Ivan Jedvaj (Zagreb school of economics and management)
Dr. Wahidin Wahab (Universitas Indonesia - Indonesia)
Prof. Ahmad Binadja (Universitas Dian Nuswantoro - Indonesia)
Prof. Dr. Ir. Riri Fitri Sari. M.Sc. MM (Universitas Indonesia - Indonesia)
Prof. Dr. Ir. Harry Sudibyo, M.Sc (Universitas Indonesia - Indonesia)

Technical Program Committee :

Dr. Guruh Fajar Sidiq, S.Kom, M.Cs (Chair)
Dr. M. Ary Heryanto (Universitas Dian Nuswantoro - Indonesia)
Dr. Heru Agus Santoso (Universitas Dian Nuswantoro - Indonesia)
Dr. Dian Retno Sawitri (Universitas Dian Nuswantoro - Indonesia)
Dr. Herwin Suprijono (Universitas Dian Nuswantoro - Indonesia)
Dr. Hector Sanches Lopez (Universitas Dian Nuswantoro– Indonesia)
Dr. Mochamad Facta (Universitas Diponegoro - Indonesia)
Dr. Supari (Universitas Semarang - Indonesia)
Dr. Mohamad Haddin (Universitas Islam Sultan Agung - Indonesia)

Dr. Dimas Anton Asfani (Institut Teknologi Sepuluh November - Indonesia)
Dr. Djoko Purwanto (Institut Teknologi Sepuluh November - Indonesia)
Dr. Sarjiya (Universitas Gajah Mada - Indonesia)
Dr. A.N. Afandi (Universitas Negeri Malang - Indonesia)
Dr. Danang Wijaya (Universitas Gajah Mada - Indonesia)
Dr. I Wayan Mustika (Universitas Gajah Mada – Indonesia),
Dr. Jumanto, M.Pd (Universitas Dian Nuswantoro - Indonesia)
Dr. Tari Eszter (Universitas Dian Nuswantoro - Indonesia)
Dr. Wahyudi (UNniversitas Diponegoro - Indonesia)
Dr. Eng Suharyanto (Universitas Gajah Mada - Indonesia)
Dr. Paulus Insap Santoso (Universitas Gajah Mada - Indonesia)
Dr. Lukito Edi Nugroho (Universitas Gajah Mada - Indonesia)
Dr. Kusprasapta Mutijarsa (ITB - Indonesia)
Dr. Tole Sutikno (Universitas Ahmad Dahlan - Indonesia)
Dr. Arif Muntasa (Universitas Trunojoyo - Indonesia)
Dr. Ketut Edi Purnama (Institut Teknologi Sepuluh November – Indonesia)
Dr. Darlis Herumurti (Institut Teknologi Sepuluh November - Indonesia)
Dr. I.G.P Asto Budi Tjahyanto (Universitas Negeri Surabaya - Indonesia)
Dr. Lie Jasa (Universitas Udayana - indonesia)
Dr. Munawar Riyadi (Universitas Diponegoro - Indonesia)
Dr. Feri Yusivar (Universitas Indonesia - Indonesia)
Dr. Riko Arlando Saragih, S.T., M.T. (Universitas Kristen Maranatha - Indonesia)
Dr. I Gede Pasek Suta Wijaya (Universitas Mataram - Indonesia)
Dr. Saiful Bukhori (Universitas Jember - Indonesia)
Dr. Moh Khairudin (Universitas Negeri Yogyakarta - Indonesia)
Dr. Eng. Yuliman Purwanto (Universitas Dian Nuswantoro - Indonesia)
Dr. Ruri Suko Basuki (Universitas Dian Nuswantoro - Indonesia)
Dr. Arief Sulaiman (Universitas Dian Nuswantoro -Indonesia)
Dr. Muljono (Universitas Dian Nuswantoro - Indonesia)
Dr. MG Yuantari (Universitas Dian Nuswantoro - Indonesia)
Dr. Eng. Arief Udhiarto, S.T, M.T (Universitas Indonesia - Indonesia)
Dr.Ir. Aries Subiantoro, M.SEE (Universitas Indonesia - Indonesia)
Dr. Hamdani (Universitas Mulawarman - Indonesia)

GENERAL INFORMATION

VENUE

Conference Venue

iSemantic 2018 will be held in Universitas Dian Nuswantoro, Semarang, Indonesia. It is placed on Building E (Gedung E) 3rd floor of the University Campus.

All sessions will take place in the Faculty of Computer Science that are at the fifth (5th) Floor of Building D (Gedung D).

It is easily accessed by both public transportation and personal transportation. It takes around 12 minutes from Ahmad Yani Airport.

Located in a 4.5 Ha area, the campus is surrounded by office, education, commercial building and housing area.

Address

Building D (Gedung D), 1st Floor, Kampus II
Lembaga Penelitian dan Pengabdian Masyarakat
Universitas Dian Nuswantoro

Jalan Nakula 1 No. 5-11, Semarang Indonesia 50131

http://isemantic.dinus.ac.id/2018/

MAPS



KEYNOTE SPEAKER 1

Prof. Dr. Nurul Taufiqu Rochman, M.Eng.

Indonesian Institute of Sciences



Biography

Nurul Taufiqu Rochman is a Senior Physic Scientist in Indonesian Institute of Sciences. He is known as nanotechnology expert in Indonesia. Many publications and patents have been granted by him. High Milling 3D Motion nanoparticle is one sample of his achievement. His interest in nano technology makes impact in some aspects of human life such as ink nano that contains no chemical material. He also contributes in nano center development in Indonesia.

KEYNOTE SPEAKER 2

Prof. Peter Wushou Chang, MD, MPH, ScD

Tufts University School of Medicine



Biography

Prof. Peter Wushou Chang, MD, MPH, ScD is the Vice President of International Health Literacy Association. He received his Sc.D in Laboratory of Radiobiology, Department of Cancer Biology, Harvard University, School of Public Health in 1992. He is also Advisory Committee member, Low carbon and green house gas reduction Strategies; Environmental Protection Administration (EPA) from Jan. 2015. He was former Dean for International Affairs, International Office, Taipei Medical University. He received Research Award, National Science Council, Taiwan, 1996, 1997, 1998, 1999, 2000, including Candidate of Outstanding Research Award in 1998, 2000, 2001 and Award winner, 2017 International Health Literacy Award, Institute for Healthcare Advancement, IHA, Irving, California, USA.

KEYNOTE SPEAKER 3

Dr. Reece Hinchcliff

University of Technology Sydney



Biography

Dr. Reece Hinchcliff the Acting Discipline Lead and Director of Studies in Health Services Management at the University of Technology Sydney (UTS). His work examines how regulatory schemes and innovative analytical techniques can be applied to improve the quality and safety of health services and systems. His research has been published in 28 peer-reviewed journal papers that have been cited on over 550 occasions, producing a Google Scholar i10-index of 17 and H-index of 14. He has also presented at international conferences on more than 20 occasions, including twice as a keynote speaker. He hold concurrent appointments at several universities, including as a Conjoint Senior Lecturer at the University of New South Wales (UNSW) and an Honorary Senior Research Fellow at the Australian Institute of Health Innovation, Macquarie University. He have also been an editorial board member of BMC Health Services Research since 2015 and served on two government policy committees.



PROCEEDINGS

List of Papers

PAPER TITLE	PAGES
Service-oriented Business Intelligence (SoBI) for Academic and Financial Data Integration in University	1-5
Evaluation of Farmers Terms of Trade Indonesian Agriculture Forecasting Method	6-10
Improving Information Service Performance of Family Businesses through IT Governance	11-16
Measuring Maturity Index of Risk Management for IT-Governance Using Fuzzy Ahp and Fuzzy Topsis	17-22
Analysis of Potencies and Investment Opportunities in East Java Based on Geographic Information System	23-28
Time and cost optimization of business process distribution company using goal programming	29-34
Time and Cost Optimization using Goal Programming and Priority Scheduling	35-40
Vendor Track Record Selection Using Best Worst Method	41-48
Strategy for Optimizing Cost and Time of After Sales using Integer Programming	49-54
Improving Business Process by Evaluating Enterprise Sustainability Indicators using Fuzzy Rule Based Classification	55-60
IoT Communication System using Publish-Subscribe	61-65
Investigation of Ferroresonance Physical Behaviours on Three Phases Transformer with Unsymmetrical Core Leg	66-70
Autonomous Controller for MetOcean Monitoring Station	71-76
Optimization Service Discovery in Wireless Balloon Network	77-82
Evaluation of Normalization in Fake Fingerprint Detection with Heterogeneous Sensor	83-86
Pulse Width Modulation (PWM) and Pulse Amplitude Modulation (PAM) Technique for Medium-Speed BLDCM in Electric Vehicle Application	87-92
Signal Measurement Consistency of ECVT Data Acquisition System using Capacitor Array	93-97
Control System Scheme for Shuttlecock Thrower Robotic Based on PID-CSA	98-103
A Natural Phenomenon Emphasizing of the Thunderstorm Processes and Salmon Migration for the Optimal Power Production of the Power System Operation	104-109

Performance Evaluation of Speed Controller Permanent DC Motor in Electric bike Using Fuzzy Logic Control System	110-115
Internet of Things on monitoring and control system in server area	116-120
Soft Starting & Performance Evaluation of PI Speed Controller for Brushless DC Motor Using Three Phase Six Step Inverter	121-126
Performance Robustness of PID Controller in Buck Converter For Cooling System	127-132
Analysis of 4G Network and Chat Applications to Smartphone Battery Life	133-138
The Performance of FOSMC and Boundary - SMC in Speed Controller and Current Regulator for IFOC-Based Induction Motor Drive	139-144
Comparison of Perturb and Observe, First Order Differential Algorithm and Newton Raphson Methods for PV Application in DC Microgrid Isolated System	145-150
Design and Realization of Low-Cost Wireless Remote Infusion Monitoring System	151-154
Speed Identification As Vehicle Safety Feature Based Eulerian Color Magnification	155-158
Reverse Polarity Scanning Method on 64-Channel Capacitive Sensor to Improve the Performance of ECVT System	159-165
Comparison of Cascade and Feedforward-Feedback Controllers for Temperature Control on Stirred Tank Heater Systems	166-170
Detecting Business Process Anomaly Using Graph Similarity Based on Dice Coefficient, Vertex Ranking and Spearman Method	171-176
Job Performance Analysis Using Factor Evaluation System and Process Mining	177-183
Media Economics and the Use of Technology in the Production of Local Gaming Industry (A Study at Creacle and Wisageni Game Studio Yogyakarta)	184-189
Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifier	190-195
Interaction Design Consideration for Senior High School Students: A Usability Evaluation of Go-Jek Mobile Applications	196-201
Prediction of Movie Sentiment based on Reviews and Score on Rotten Tomatoes using SentiWordnet	202-206
Alignment Model of Quality Assurance System of Higher Education And Performance Measurement Based on Framework CobiT 5	207-213

Classification of Governor's Public Report from SMS LaporGub Using Naive Bayes Classifier Method	214-218
A User-centered Design for Redesigning E-Government Website in Public Health Sector	219-224
Determining The Senior High School Major Using Agglomerative Hierarchial Clustering Algorithm	225-228
Influence of Word Normalization and Chi-squared Feature Selection on SVM Text Classification	229-233
Sentiment Analysis using Weighted Emoticons and SentiWordNet for Indonesian Language	234-238
Word Sense Disambiguation in Bahasa Indonesia Using SVM	239-243
Developing Word Sense Disambiguation Corpuses using Word2vec and Wu Palmer for Disambiguation	244-248
Emotion Label from ANEW dataset for Searching Best Definition from WordNet	249-252
Searching Word Definitions in WordNet Based on ANEW Emotion Labels	253-256
Digital Literacy Skills of Informatics Engineering Education Students as The Basis for Online Learning Implementation	257-260
Sentiment Analysis of Indonesian News Using Deep Learning (Case Study: TVKU Broadcast)	261-265
Measurement of Spinal Curvature for Scoliosis Classification	266-270
Mining Medication Behavior of The Completion Leprosy's Multi-Drug Therapy in Indonesia	271-274
Mobile Health Nutrition Book Design to Prevent Stunting at Childreen <5 years	275-279
Improvement Fuzzy C-Mean using Local Laplacian Filter for Image Segmentation	280-285
Performance Comparative of AODV, AOMDV and DSDV Routing Protocols in MANET Using NS2	286-289
Bit Localization in Least Significant Bit using Fuzzy C-Means	290-294
Pork and Beef Features Extractions	295-298
Gossip Algorithm Implementation for Network Protocol	299-303
A Robust Non-Blind Image Watermarking Method using 2-Level HWT-DCT	304-308
Classification Method of Academic Acquisition Value Through Student Behavior	309-312
A Comparative Study of some Video Players based on Visual Output Quality	313-317
Classification of Music Moods Based on CNN	318-321

Comparison Of sGA And SEGA Methods To Solve The Problem Of Power Generation And Power Losses On Distributed Generating Systems	322-327
The using of Object-Oriented Method to Developing Android Application for Mapping Distribution of MSMEs (Micro, Small Medium Enterprises)	328-332
RESTful Web Service Optimization with Compression and Encryption Algorithm	333-337
Performance Discovery in Ubiquitous Networks	338-344
Design and Development of REST-based Instagram Spam Detector for Indonesian Language	345-350
Parking Space Detection Using Quaternionic Local Ranking Binary Pattern	351-355
A Comparison Study of Search Strategy on Collecting Twitter Data for Drug Adverse Reaction	356-360
Ultranus: a Novel Indonesian Cultural Game using Artificial Intelligence	361-366
OutPatient Registration System Based On Android At BBKPM Surakarta	367-372
Intelligent System for Agent in Educational Game using Dynamic Gram Similarity	373-376
A Performance Comparative of Vertical Fragmentation Table using Bond Energy and Graph Based Vertical Partitioning Algorithm	377-380
QoS Improvement Analysis of VoIP Service Which Uses Overlay Network. Case Study: Calling AWS VoIP Gateway From Bandung, Indonesia	381-387
Surveillance Embedded IP Camera with Integrated Cloud Storage and Image Deblurring Using Richardson-Lucy	388-393
Supplier Selection Using Combined Method of K-Means and Intuitionistic Fuzzy Topsis	394-399
Heart Rate Monitoring System in Virtual Environment: The Case of Daily Exercise with Virtual Reality	400-405
Facial Emotion Recognition based on Viola-Jones Algorithm in the Learning Environment	406-410
Handwriting Recognition using Eccentricity and Metric Feature Extraction based on K-Nearest Neighbors	411-416
Dinus Intelligent Assistance (DINA) Chatbot for University Admission Services	417-423
Comparison Of Severity On Mobile Government Application Mobile	424-428
Prediction of Glucose Content and Level of Tomato Maturity With Backpropagation Based On Image Vector Feature	429-434

Conformance Checking Evaluation of Process Discovery Using Modified Alpha++ Miner Algorithm	435-440
Implementation of Bat Algorithm for COCOMO II Optimization	441-446
Time and Cost Optimization Using Dynamic Programming and FMS Scheduling	447-452
Comparison Of AHP and BWM Methods Based on Geographic Information System For Determining Potential Zone of Sand Mining	453-457
Time and Cost Optimization of Parallel Services on Line Balancing Problem using Integer Programming	458-462
Integration of Fuzzy C-Means Clustering and TOPSIS (FCM-TOPSIS) with Silhouette Analysis for Multi Criteria Parameter Data	463-468
Performance Time Evaluation of Domestic Container Terminal using Process Mining and PERT	469-475
An Experimental Study of Supervised Sentiment Analysis Using Gaussian Naïve Bayes	476-481
Improving The Accuracy of Naïve Bayes Algorithm for Hoax Classification Using Particle Swarm Optimization	482-487
A Comparison of Data Mining Techniques for Suicide Attempt Characteristics Mapping and Prediction	488-493
Bayesian Network Optimization for Earthquake Prediction Using Simulated Annealing	494-497
Automatic Gestational Age Estimation by Femur Length using Integral Projection from Fetal Ultrasonography	498-502
A Lexicon-based Sentiment Analysis for Amazon Web Review	503-508
CBIR for Herbs Root using Color Histogram and GLCM based on K-Nearest Neighbor	509-514
Image Enhancement Segmentation And Edge Detection In MRI For Mammogram Disease	515-521
Identification of Regional Dialects Using Mel Frequency Cepstral Coefficients (MFCCs) and Neural Network	522-527
LoG based Watermarking over Region Magnitude using Chinese Remainder Theorem	528-532
The Determination of Cluster Number at K-Mean Using Elbow Method and Purity Evaluation on Headline News	533-538
Survey on pedestrian-dynamics models for evacuation process based on game theory	539-544
Action Training and Assessment System with Kinect	545-548
Customer Behavior Analysis Using Data Mining Techniques	549-554

Face Detection Using Haar Cascade in Difference Illumination	555-559
Search Engine for Kids with Document Filtering and Ranking Using Naive Bayes Classifier	560-564
Automatic Essay Grading System for Japanese Language Examination Using Winnowing Algorithm	565-569
Implementation Of Neural Network Backpropagation Using Audio Feature Extraction For Classification Of Gamelan Notes	570-574
Optimization of Evaluation and Monitoring Activities on Student Industrial Practices through Web-based Information System	575-579
Implementation of Weighted Naive Bayes Algorithm for Major Determination in Indonesian High School	580-584
Heart Beat Based Drowsiness Detection System for Driver	585-590
Economic Studies of the Wind Turbin-Diesel Hybrid Power Generation System (case study at: Queen of the South Beach Resort Hotel, Yogyakarta, Indonesia)	591-596
Classification of cataract Slit-lamp image based on machine learning	597-602
Improving Classifier Performance Using Particle Swarm Optimization on Heart Disease Detection	603-608
Improved Edge Detection Based on Adaptive Gaussian Smoothing in X-Ray Image	609-616

Service-oriented Business Intelligence (SoBI) for Academic and Financial Data Integration in University

Ramos Somya
Faculty of Information Technology
Satya Wacana Christian University
Salatiga, Indonesia
ramos.somya@uksw.edu

Danny Manongga
Faculty of Information Technology
Satya Wacana Christian University
Salatiga, Indonesia
danny.manongga@uksw.edu

Magdalena. A. Ineke Pakereng
Faculty of Information Technology
Satya Wacana Christian University
Salatiga, Indonesia
ineke.pakereng@uksw.edu

***Abstract*— Previously, data integration in Business Intelligence (BI) is done by ETL process (Extract, Transform, and Loading). Service-oriented Business Intelligence (SoBI) offers data integration process in a simpler way when compared to the ETL process in BI. In this study, the SoBI will be implemented to integrate academic and financial data at Satya Wacana Christian University (SWCU) into Data Warehouse. The Data Warehouse was developed with Galaxy scheme and implemented with SQL Server 2012. The results of the study show that to apply SoBI, dashboard application must be made that serves to manage data integration process. Data integration on SoBi is done with the help of Web Service by building Service Provider at location of data source and Service Requestor on dashboard application. The Service Provider can be called by the dashboard application to perform the data retrieval process from the data source and transfer data into the Data Warehouse, so that the ETL process is done entirely by the Web Service. In addition, the dashboard application also serves to synchronize data from data sources through Web Service so that the data to be integrated is also more consistent.**

Keywords—SoBI, Business Intelligence, Data Warehouse, Data Integration

I. INTRODUCTION

Business Intelligence (BI) is a concept about data integration and data processing from various sources for better decision making in a company. It provides facilities for company to process data into information and knowledge. Decision makers use the information and knowlegde to make better business strategies [1]. BI requires a Data Warehouse that is used to store all data derived from different data sources. In general of BI concepts, data from these sources will be integrated into the Data Warehouse through the ETL (Extract, Transform, and Loading) process.

ETL is the most important process in Data Warehouse. It has responsibility in the quality of data integration and it can be the most complicated stage in Business Intelligence [2]. In this ETL process, there are data retrieval from data source (extract), data transformation process (transform), and the process of moving all data into Data Warehouse (loading). Stages in this ETL take a relatively long time, especially at the transform stage. In this transform stage, the developer must change all data originating from the data source into a certain format. The data cleansing process should also be done to maintain the accuracy of the data before all data is transferred into the Data Warehouse.

Along with the development of information technology, emerged a new concept in BI called SoBI (Service-oriented Business Intelligence). This concept offers a data integration process in a simpler manner when compared to the ETL process on the BI concept in general [3].

In this study, the SoBI concept will be applied to integrate data at Satya Wacana Christian University (SWCU). The data generated by the operations at SWCU is very diverse, ranging from academic data, research data, library data to financial data. The process of data integration is necessary because the data is spread in various sources and need to be integrated in Data Warehouse to facilitate the university in accessing data for decision making purposes.

This study will examine the process of integration of academic data and financial data at SWCU with the concept of SoBI. The academic data will be taken from Academic Information System (SIASAT) database of Satya Wacana Christian University. While financial data will be taken from Satya Wacana Financial System database (SIKASA). The data will be moved into the Data Warehouse so that the academic and financial department of SWCU can access the data easily for decision making purposes. Data Warehouse was developed using Galaxy schema and SQL Server 2012. Galaxy Schema is used in this research because Data Warehouse is designed to provide 2 fact tables, they are a fact table for academic subject and finalcial subject. There are 3 schemas in Data Warehouse concept, and the only schema that allow the user to use multiple fact tables is only Galaxy Schema. Dashboard application in the form of Online Analytical Processing (OLAP) will also be developed to present data in Data Warehouse into information that is easy to access.

II. LITERATURE REVIEW

A. Previous Research

Research on data integration has been done before by utilizing tools of Pentaho Data Integration and Oracle Data Integrator. In this study, data extracted from the data source is converted into a spreadsheet format before it is integrated. The data that has been integrated can then be processed into information for analysis purposes [4].

Various problems encountered in the data integration process have also been studied in order to improve the quality of the data that are integrated. This study yields 4 suggestions to give simpler method about data integration in Data

Warehouse. These suggestions are obtained after testing the data integration using SQL-Server Integration Services (SSIS) [5]. SSIS provides mechanism to import and export data automatically. User simply provide data in a certain format and then load data using the feature provided by the SSIS.

This study also examines the process of data integration. However, the approach used is different. The SoBI concept will be used to maximize the ETL process so that the process can be done simpler. The university's academic data and financial data will be integrated into Data Warehouse developed using SQL Server 2012. The position of the data is deemed in accordance with the concept of SoBI because the data is generated by the Online Transaction Processing (OLTP) application at SWCU, accessed using Web Service for data integration process

B. Service-Oriented Business Intelligence (SoBI)

SoBI is a combination of the concept of Service Orientation (SO) and Business Intelligence (BI) which seeks to solve the problem of separate system integration within a company [6]. This concept tries to provide a simpler data transformation mechanism. Objectives of SoBI include: providing a good system implementation framework, integrating the system at the appropriate level of architecture and providing a simple implementation for data transformation. Fig. 1 is a view between SO and BI.

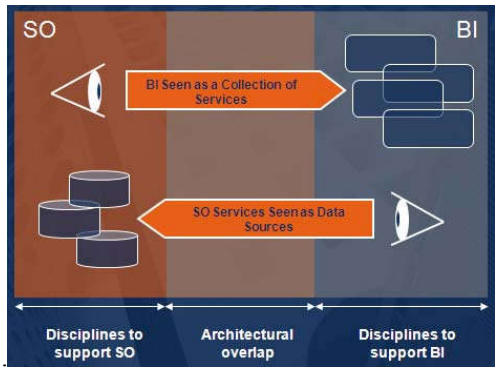


Fig 1. The views of SO and BI [6]

From SO perspective, BI is seen as a collection of services. From BI perspective, SO is seen as a collection of data sources. SoBI architecture allows BI data in the Data Warehouse is available as a service for other applications in the architecture. With this architecture, the application will get a simple way to access data for BI purposes. So the BI architecture will be an integrated component with SO.

C. Business Intelligence (BI)

Business Intelligence is defined as knowledge gained about a business through the use of various hardware/software technologies which enable organizations to turn data into information [7]. BI provides Executive Information System (EIS) for decision makers to access important information from database. They can make better business strategies based on the information derived from database. BI encompasses data integration, data processing, and decision support system in a company.

D. Data Warehouse

Data Warehouse is the center of Business Intelligence. It stores the data in a centralized place in order to give simple access for decision making puspose. Online Analytical Processing (OLAP) application can be used to access and process the data into information.

In physical implementation, Data Warehouse is a database, but it has different concept in method design. In Relational Database, the designer uses Entity Relationship Diagram (ERD) to design the database. Meanwhile, Data Warehouse uses schemas in the design process [8].

There are 3 schemas in Data Warehouse, they are Star Schema, Snowflake Schema and Galaxy Schema. In this research, Galaxy Schema will be used to store the data in Data Warehouse. Galaxy Schema is seen as several Star Schemas in one schema. It consists of more than one fact tables and they share dimension tables [9].

E. Web Service

Web Services is a software application that can be accessed remotely using a different language based on XML (eXtensible Markup Language) [10]. Web Services technology offers the ease of bridging information resources without questioning the difference in technology used in each source. For example, an information site is built using Oracle database, while other sites are built using MySQL, therefore, Web Services application will overcome this difference. Web Services is necessary because nowadays, ranging from the hardware, operating systems, applications and programming languages are increasingly diverse.

III. SYSTEM PLANNING

The process of academic and financial data integration is done with the concept of SoBI. In this concept, the data source will be identified for the data retrieval process. Fig. 2 is data integration architecture.

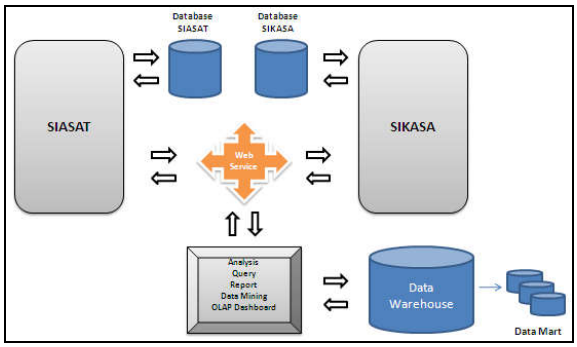


Fig. 2. SoBI data integration architecture

Fig. 2 shows the data integration architecture using SoBI. The data source is taken from academic data in database of SIASAT and financial data located in SIKASA database. The process of data retrieval is done using Web Service technology, where Service Provider is embedded in SIASAT and SIKASA applications. Dashboard application served as Service Requestor that will call the Service Provider in SIASAT and SIKASA to send the requested data. The requested data will be

sent in XML format via Web Service and processed by dashboard application which to be stored in Data Warehouse.

Data Warehouse is designed using Galaxy Schema approach and implemented using SQL Server 2012. In this Galaxy scheme, there are several fact tables that are shared by several dimension tables. Fig. 3 is the created Galaxy scheme.

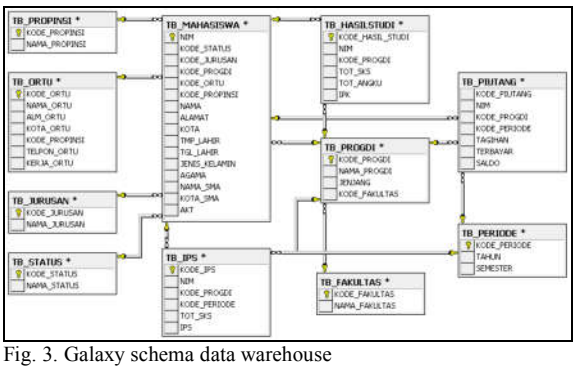


Fig. 3. Galaxy schema data warehouse

Fig. 3 shows a Data Warehouse design that is created using Galaxy Schema. In this schema there are 4 (four) fact tables, namely TB_MAHASISSWA, TB_IPS, TB_HASILSTUDI and TB_PIUTANG.

These four tables of facts share dimension tables, i.e. TB_IPS and TB_PIUTANG tables share TB_PERIODE table dimensions. While, TB_MAHASISWA, TB_IPS, TB_HASILSTUDI and TB_PIUTANG share the TB_PROGDI table dimensions.

IV. EXPERIMENTAL RESULT

The following will present the results of the implementation of the Service-oriented Business Intelligence (SoBI) design that has been made in the previous section. Based on the design, the steps that must be done are:

1) Identifying data to be retrieved from the data source: The first step is to determine type of data that will be taken from the data source from SIASAT and SIKASA databases. These data are tailored to the Data Warehouse design that was previously designed. In this experiment, the data to be retrieved are:

a) Student data: NIM (student registration number), status code, department code, study code, parent code, province code, name, address, city, place of birth, date of birth, gender, religion, name of high school, location of high school and class of high school study.

b) Semester Achievement Index data (IPS): IPS code, NIM, study program code, period code, total SKS (credit hours) and IPS.

c) Cumulative Achievement Index data: GPA code, NIM, study program code, total credits, cumulative total number and GPA.

d) Student Account Data: receivable code, NIM, study program code, period code, bill number, paid amount and balance amount.

2) Create a Web Service to retrieve data: after data to be taken is identified, the second step is to create a Web Service in the form of Service Provider on SIASAT and SIKASA that serves to retrieve the data. The implementation of the Service Provider are: getDataHasilStudi(), getDataIPS(), getDataMahasiswa(), getKodeProgdi(), getKodeProgdiMhs(), and getPeriodeIPS(). Overall, the Web Service on SIASAT can be seen in the Class Diagram resulting from Microsoft Visual Studio 2010 in Fig. 4.

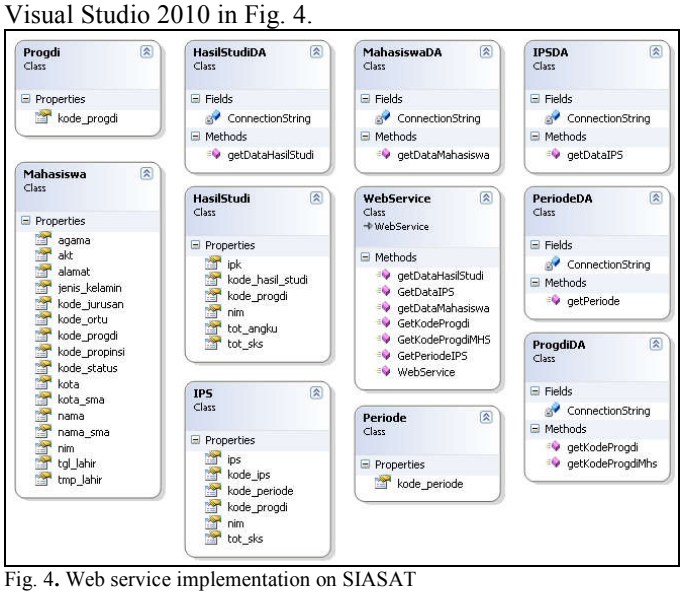


Fig. 4. Web service implementation on SIASAT

Meanwhile, the implementation of Web Service (Service Provider) on SIKASA is getDataPiutang () and getPeriodePiutang. This Web Service is illustrated in Fig. 5.

```
require_once('nusoap/nusoap.php');
$server = new soap_server;
$server->register('getDataPiutang');
$server->register('getPeriodePiutang');
$HTTP_RAW_POST_DATA = isset($HTTP_RAW_POST_DATA) ? $HTTP_RAW_POST_DATA : '';
$server->service($HTTP_RAW_POST_DATA);
```

Fig. 5. Web service implementation on SIKASA

Web service on SIKASA is created using NuSOAP library because SIKASA is a web application developed using PHP programming language.

3) Calling the Web Service: The third step is to create a dashboard application that is used as a Service Requestor, which is in charge of calling a Web Service (Service Provider) that has been made before. With this application, the process of taking and loading data into the Data Warehouse will be done from this application. Fig. 6 is a view for the dashboard application.



Fig. 6. Page for data loading

4) Performing data loading: The last step is to load data into the Data Warehouse using the dashboard application that has been created. In this experiment, loading student receivable data will be made. This process is shown in Fig. 7.

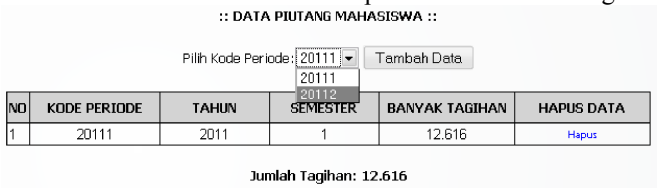


Fig. 7. Process of added receivable data

Fig. 7 is the process of added receivable data into the Data Warehouse. To add data, the user must select the period code via combo box. The period code in the combo box is taken from the SIKASA database via the Web Service. After selecting the period code, the data adding process can be done by pressing the Add Data button. The results are shown in Fig. 8.

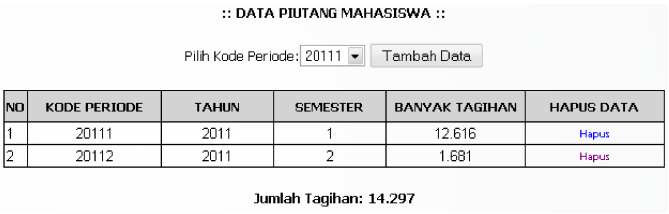


Fig. 8. Result of data loading

The dashboard application in SoBI can be equipped with features to manage data in the Data Warehouse more easily. This feature can be used to monitor data from data sources. Thus, if at the data source location there is data that has not been included in the Data Warehouse, then the system can display this information. This is shown in Fig. 9.



Fig. 9. IPS Data information on dashboard

The SoBI dashboard application in Fig. 9 can be used to easily monitor existing data in the data source. This is evidenced by the display of the latest IPS data from data sources captured and presented in combo box in Fig. 9. To add this data, the user only needs to select the period code in the combo box and press the Add Data button. Automatically, data from the data source will be retrieved through the Web Service and will be entered into the Data Warehouse.

After all data is integrated in the Data Warehouse, data can be accessed and processed into information. Information can be displayed on the dashboard application for the analysis process.

V. CONCLUSION

Based on the research that has been done, it can be concluded that by using the concept of SoBI, the process of integration of academic and financial data from SIASAT and SIKASA can be done more easily and faster when compared to using the concept of BI in general. This can be seen when the BI concept is used, data integration process must be through a complicated ETL (Extract, Transform, and Loading) process, whereas when SoBI concept is used, data integration can be done more simply by using Web Service.

Data integration using the SoBI concept requires a dashboard application that is used to manage data integration processes from a data source. The data in the data source should also be stored in the database in order to be accessible by Web Service. The dashboard application on SoBI can be equipped with features to manage data from data sources. Thus, the user can see the latest data from the data source and can put it into the Data Warehouse more easily through the use of Web Service. Data processing using Data Mining can be done for the future work of this research.

REFERENCES

[1] M. Prabhavathy, and K. Sivasankari, "Federated Query Processing Service in Service Oriented Business Intelligence," International Conference on Engineering Technology and Science, vol. 3, pp. 1266-1272, February 2014.

[2] A. Karagiannis, P. Vassiliadis, and A. Simitsis, “Scheduling strategies for efficient ETL execution,” Inf. Syst., vol. 38, no. 6, pp. 927–945, September 2013.

[3] A. Karagiannis, P. Vassiliadis, and A. Simitsis, “Scheduling strategies for efficient ETL execution,” Inf. Syst., vol. 38, no. 6, pp. 927–945, September 2013.
[4] A. M. I. Florea, V. Diaconita, and R. Bologa, "Data Integration Approaches using ETL," Database System Journal, vol. 6, no. 3, pp. 19-27, 2015.
[5] S. Shivtare, and P. Shelar, "Data Warehouse with Data Integration: Problems and Solution," IOSR Journal of Computer Engineering (IOSR-JCE), pp. 67-71, 2015.
[6] Gordon, Sean, and Grigg, Rob, Service-oriented Business Intelligence (SoBI), Microsoft, 2005.
[7] V. Leat, "Introduction to Business Intelligence," IBM Software Group, 2007.
[8] Oracle, "Data Warehousing Concepts," Oracle9i Data Warehousing Guide Release 2 (9.2), 2002.
[9] G. Saxena, and B. B. Agarwal, "Data Warehouse Designing: Dimensional Modelling and E-R Modelling," International Journal of Engineering Inventions, pp. 28-34, 2014.
[10] F. N. Allokendek, J. Soetikno, and A. Ashari, “Integrasi Database DISDUKCAPIL dan Database KPU Kabupaten Maros Memanfaatkan Web Services,“ IJCCS (Indonesian J. Comput. Cybern. Syst.), vol. 7, no. 1, pp. 1-12, 2013.

Evaluation Forecasting Method of Farmers Terms of Trade Indonesian Agriculture

Ifran Lindu Mahargya, Rizchi Eka Wahyuni, Guruh Fajar Shidik
Faculty of Computer Science
University Dian Nuswantoro
Semarang City, Indonesia
ifran.mahargya@gmail.com, rizchi.ew@gmail.com, guruh.fajar@research.dinus.ac.id

***Abstract*—The agricultural sector is still the most highly employed sector. One indicator that can be used to measure success in the agricultural sector is Farmers Terms of Trade that is usually called *Nilai Tukar Petani* (NTP) in Indonesia. NTP is an indicator that can be used to measure purchasing power of farmers. NTP data is currently released at the beginning of every month based on the previous month's data condition. If NTP data can be predicted earlier, the government can take the right policy in advance. Some methods were used for forecasting data. In this research, we proposed NN, SVM and Multiple Linear Regression methods to forecast NTP. We conclude that Multiple Linear Regression is the best method to forecast NTP with RMSE 0.0009.**

Keywords— Farmers Terms of Trade, Neural Network, Support Vector Machine, Linear Regression.

I. INTRODUCTION

The agricultural sector is the most labor intensive sector in Indonesia, where approximately 31.86% of labor is absorbed by the sector [1]. This reflects that the agricultural sector is still a cornerstone for the Indonesian population that contribute to Gross Domestic Product (GDP) of 13.45% at 2016 [1]. Accordingly, an indicator is needed to accurately measure the purchasing power of farmers as one of the main actors of the agricultural sector. An indicator to measure the purchasing power of farmers in a region is Farmers Term of Trade that is usually called "*Nilai Tukar Petani"* (NTP) in Indonesia. NTP is the ratio of the price index of agricultural commodities produced by farmers (It) to the price index of goods or services paid by farmers for household consumption and production costs (Ib). The price index received by farmers (It) is a price index to show the development of producer prices on farmers' production. The index can show fluctuations in the prices of goods produced by farmers. In addition this index is used as supporting data in calculating agricultural sector revenues. The price index received by farmers (It) is calculated based on the selling value of agricultural products produced by farmers, including the sectors of rice, pulses, livestock products, plantations of people, vegetables, fruits, and fishery products (capture fisheries and cultivation). The price index paid by farmers (Ib) is a price index to show the development of prices of farmer's household needs, both the need for household consumption and the need for agricultural production processes. The index can show fluctuations in the prices of goods consumed by farmers which are the largest part of rural communities, as well as fluctuations in the prices of goods needed to produce agricultural products. The development of the price index paid by farmers (Ib) can illustrate the development of inflation in rural areas. The price index paid by farmers (Ib) is calculated based on the price index that must be paid by farmers in meeting the needs of life and the addition of capital goods and production costs which are further divided into food, non-food goods and services.

In general Farmers Term of Trade generates 3 meanings:

1. NTP >100 indicates that NTP in a certain period is better than the NTP of the base year or it can be said that the farmer has a surplus.
2. NTP = 100 shows that the NTP in a certain period is the same as the NTP for the base year or it can be said that the farmer has broken even.
3. NTP <100 shows that NTP in a certain period decreases compared to the base year NTP or it can be said that the farmer has a deficit.

NTP is provided by Indonesian Central Bureau of Statistics (BPS) and released at the beginning of each month. NTP Score has national and regional analysis unit (province) according to each region, NTP score used research that has been done is NTP score for Central Java Province. In this research we did forecasting the NTP score in Jawa Tengah Province. Jawa Tengah Province as one of the provinces that has an agricultural potential and need an appropriate policies implementation in the agricultural sector. By receiving the NTP score earlier, the government is expected to provide prevention policies for the bad NTP condition.

Some methods have been used for forecasting in agriculture sector. Pei Wang [2] conducted a study to forecast China Farmers' Earnings using Neural Network (NN). The study has concluded that using NN in predicting earnings of Chinese farmers has helped solve problems and difficulties caused by complexity in forecasting and adjusting weight. In addition, the research has helped to improve the accuracy of forecasting. Dianhui Mao, et.al [3] using Support Vector Machines (SVM)

to predict soil erosion. Betty J. Sitienei, et.al. [4] use regression models to predict tea crop yield responses to climate change.

This research evaluate classification accuracy of the Farmers Terms of Trade Indonesian Agriculture. There are SVM, NN and Linear Regression following the existing work such as Muniz et. al. [5] and Gupta [6] research.

II. FORCESTING METHODS

A. Neural Network

Neural Network (NN) is a method that combined several simple processing to parallel distributed processing that has a natural tendency to learning from the past data and store it to the knowledge and can be used when needed [7]. NN can perform recognition of past data-driven activities. The data is studied by NN so that it can choose a decision for data that has not been researched or future data. [8]. The relationship of one neuron to another neuron has a price associated with weight. Each neuron has a value associated as a neuron activation value [8]. The relationship between neurons or commonly referred to as network architecture. The neurons are collected in layers called layer neurons. The NN layer is divided into three namely the input, hidden and output layer [9].

B. Support Vector Machine

Support vector machine (SVM) is an efficient classification technique in solving nonlinear problems and introduced in 1992 by Vapnik. How SVM works is to try to find a hyperplane that maximizes the distance between classes, so SVM can ensure high generalization capabilities in future data or unseen data [10]. SVM is also used for forecasting financial time series [11]. In another research SVM is also used for prediction of hong kong horse racing [12]. Besides other related research that SVM is also used for wind turbine fault and external environmental factors, so far the existing system can only correct the error and diagnosis of wind turbine but can't predicting fault [13].

SVM identifies with low dimensional mapping to high dimensional spaces using kernel functions. Non-linear classification problem solved by SVM when the input space in the low-dimensional space has data that can't be classified, so that the higher dimension space is required using kernel function. Problems solved by SVM more often use the Nonlinear Programming Kernel function. [14].

C. Multiple Linear Regression

The model for linear additives in connecting the dependent variable with the independent variable p is as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon \qquad (1)$$

The symbol i denotes the observational data from the dependent variable Y and independent variable i to p. If the data size is defined by the symbol n, then i = 1, ..., n. So the parameters are number of p + 1 where β_0 where j = 1, ..., p is added with β_0 commonly referred to as intercept and ϵ is error.

In the least square model, the best model is obtained by calculating the minimum error or deviation between the actual data and the result data from the model. The deviation is then squared and then summarized so that no positive and negative values are considered. The values of β_0, β_1, ... βp are usually calculated with statistical or data mining software.

The value of the model $\beta 0 + \beta 1 xi1 + \ldots + \beta pxip$ is denoted by i and the error is denoted by i which can be computed by yi-yi. The number of errors should be close to 0. Variants can be estimated by the following formula:

$$s^2 = \frac{\sum e_i^2}{n - p - 1} \qquad (2)$$

The formula is also referred to as a mean-squared error (MSE). The predicted standard error is the square root of MSE [15].

D. Root Mean Square Error (RMSE)

RMSE is a value used to measure the difference between the model prediction value and the actual value. The RMSE is formulated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (X_{obs,i} - X_{model,i})^2}{n}} \qquad (3)$$

where Xobs is the value of the observation and Xmodel is the predicted value. The smaller the RMSE value of the model the better.

III. RESEARCH METHOD

In this research, we proposed step in Figure 1 to forecast the Farmers Terms of Trade.

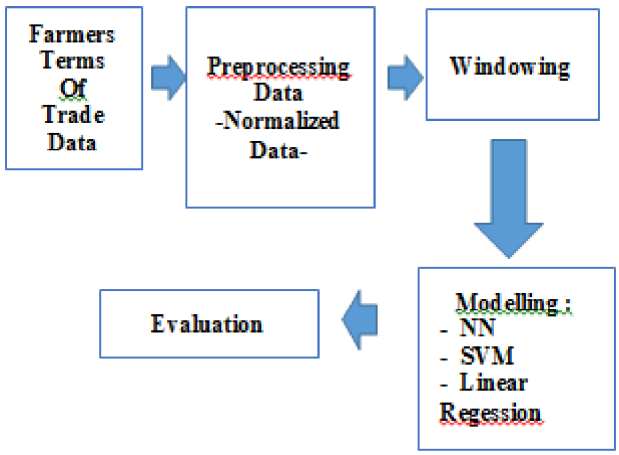


Figure 1. Step of Research Method.

A. Data

The dataset used in this study from monthly NTP data published by BPS Jawa Tengah Province consisting of Farmers Terms of Trade data from January 2008 until August 2017 [1]. NTP score is released at the beginning of month in Press Release held by Statistics Jawa Tengah Province. Then, the data released was published at officially site of Statistics Jawa Tengah Province at ARC of Press Release Menu. To get the data series, we access the Dynamic Table Menu. The Data Examples are shown in Table I.

TABLE I
SAMPLE DATASET

Month	2008	2009	2010	2011	2012	2013
JAN	106.69	98.27	100.62	102.92	106.56	106.45
FEB	105.41	98.38	100.23	103.58	105.42	105.7
MAR	103.17	98	100.22	102.84	104.51	104.59
APR	96.74	97.84	100.25	103.65	104.51	104.56
MEI	98.32	97.86	100.55	104.18	104.4	105.23
JUN	99.77	98.04	101.09	104.76	104.54	105.9
JUL	100.32	98.08	102.26	105.38	104.88	105.62
AGU	101.08	98.7	102.65	105.86	105.45	106.15
SEP	102.27	99.69	102.6	105.89	105.57	106.44
OKT	102.35	99.21	102.62	106.15	106.15	107.37
NOV	101.65	99.81	103.16	106.14	105.78	106.73
DES	102.7	100.03	103.12	106.62	106.37	101.66

B. Preprocessing

In the forecasting process using backpropagation, prior to the training process, input data and output targets should be normalized first. Normalization is scaling of values into a range of activation functions used in the network. For the binary sigmoid activation function, the data must be transformed to its output range, 0 to 1. However, it is better transformed into 0.1 to 0.9. The sigmoid function is an asymptotic function with the score not reached 0 or 1 [10]. Examples of normalization results as in Table II.

TABLE II
EXAMPLE OF DATASET NORMALIZED

Month	NTP 2016	NTP 2016 (normalized)
Jan	101.52	0.10152
Feb	100.53	0.10053
March	99.4	0.0994
April	98.99	0.09899
Mey	99.86	0.09986
June	99.64	0.09964
July	99.93	0.09993
August	100.43	0.10043
Sept	100.88	0.10088
Oct	100.15	0.10015
Nov	95.55	0.09555
Dec	99.35	0.09935

C. Windowing

Windowing is take any time series data and transform it into a "cross-sectional" format. In this research, NTP data is windowed before processed with some method. The exampled show in Table III.

D. Modelling

As explained before, in our experiment we compare the prediction techniques using NN, SVM and Linear Regression. The prediction is evaluated using Root Mean Square Error (RMSE). First we run the data using NN with 0.0005 learning rate, 1 hidden layer, 10.000 training cycle and 0.2 momentum. Second, we run the data using SVM with kernel cache 200, kernel type DOT, c 1.5 , convergence 0.001, max iteration 100.000. The last using Linear Regression without any feature selection technique. All of method we applied cross validation 90:10.

E. Evaluation

The model generated at step D will be evaluated by calculating RMSE. The predictable scenario of the prediction process is performed to produce the smallest error. The smaller the RMSE, the better the resulting model.

IV. EXPERIMENTAL RESULT

The data used in this research is data sourced from the Central Bureau of Statistics (BPS) Indonesia, especially in the area of Central Java Province. So it has good eligible. BPS is the Government of Indonesia which has authority in data and statistical activities in Indonesia. Data that has been issued by BPS is the official data. The province of Central Java is one of the agricultural regions of Indonesia that produces the largest agricultural production. Thus the region of Central Java Province becomes the area that we use in research.

The line graph of actual and prediction using three method is shown in Firgure 2. As the evaluation, we used RMSE that shown below:

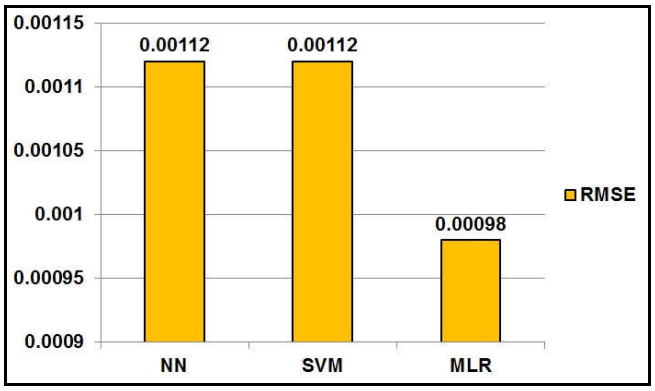


Figure 3. RMSE of three method.

Figure 3 shows that RMSE of three method is not too different. If we round it two values after comma, it is 0.001. But if we compare the RMSE of MLR is lower than others. It is indicated that for this case, MLR is better than NN or SVM.



Figure 2. The line graph of actual and prediction using three method (see online version for colours).

TABLE III
EXAMPLE OF WINDOWED DATASET

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	y
0.10669	0.10541	0.10317	0.09674	0.09832	0.09977	0.10032	0.10108	0.10227	0.10235	0.10165	0.1027	0.09827
0.10541	0.10317	0.09674	0.09832	0.09977	0.10032	0.10108	0.10227	0.10235	0.10165	0.1027	0.09827	0.09838
0.10317	0.09674	0.09832	0.09977	0.10032	0.10108	0.10227	0.10235	0.10165	0.1027	0.09827	0.09838	0.098
0.09674	0.09832	0.09977	0.10032	0.10108	0.10227	0.10235	0.10165	0.1027	0.09827	0.09838	0.098	0.09784
0.09832	0.09977	0.10032	0.10108	0.10227	0.10235	0.10165	0.1027	0.09827	0.09838	0.098	0.09784	0.09786
0.09977	0.10032	0.10108	0.10227	0.10235	0.10165	0.1027	0.09827	0.09838	0.098	0.09784	0.09786	0.09804

MLR is better than other methods because the kind of data we have used in our research is linear series of data. Despite the fluctuation of the NTP Score but still in a linear range. In our research, MLR has a smaller RMSE than SVM and NN. MLR has RMSE 0.00014 better than SVM and NN. In the results of research conducted by S.I.Abba et al [16] MLR is no better than NN, that research is prediction of water model in river. The actual data on the research is a non-linear type. In the research conducted by Lelitha Vanajakshi et al [17, 18] resulted in SVM superior to NN for short-term prediction on traffic speed data and travel time data.

V. CONCLUSION

MLR is a method capable of provising better RMSE than NN and SVM on forecast the Farmers Terms of Trade Data. MLR has a smaller RMSE than SVM and NN. Using the 12 window technique, MLR has a RMSE 0.00014 better than SVM and NN. MLR is very good at making predictions in linear time series data.

However, further research is needed to obtain the optimized variables of NN and SVM techniques. For example we need to examine hidden layers, training cycles, etc. SVM also requires advanced research such as parameter development.

REFERENCES

[1] Badan Pusat Statistik Provinsi Jawa Tengah. http://jateng.bps.go.id/site/resultTab. [accesed at 4 September 2017].

[2] Wang, Pei, et.al. “Study of Chinese Income Forest Model Based on BP Neural Network”, 2014 IEEE Workshop on Electronics, Computer and Applications, 2014.

[3] Mao, Dianhui, et.al. “Support Vector Machines with PSO Algorithm for Soil Erosion Evaluation and Prediction”, Third International Conference on Natural Computation (ICNC 2007) 0-7695-2875-9/07, 2007.

[4] J, Betty, et.al . “On the Use of Regression Models to Predict Tea Crop Yield Responses to Climate Change: A Case of Nandi East, Sub-County of Nandi County, Kenya”, Climate 2017, 5, 54; doi:10.3390/cli5030054, 2017.

[5] A.M.S.Muniz, H.Liu, K.E.Lyons, R.Pahwa, W.Liu, F.F.Nobre, J.Nadal, “Comparison among probabilistic neural network, support vector machine and logistic regression for evaluating the effect of subthalamic stimulation in Parkinson disease on ground reaction force during gait”. Journal of Biomechanics 43 (2010) 720–726.

[6] Yogesh Gupta, "Selection of important features and predicting wine quality using machine learning techniques". Procedia Computer Science 125 (2018) 305–312.

[7] D. R. Navak, A. Mahapatra, and P. Mishra, "A Survey on rainfall prediction using Artificial Neural Network", Int. J. of Comput. Applicat., Vol. 72, no. 4, pp. 32-40, 2013.

[8] Hermawan, A. 2006. Jaringan Saraf Tiruan, Teori, dan Aplikasi. Penerbit Andi, Yogyakarta.

[9] Darji. Mohini P., Dabhi. Vipul K., B.Prajapati. Harshadkumar, "Rainfall Forecasting Using Neural Network: A Survey", International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India, 2015.

[10] Vapnik, V. “The Nature of Statistical Learning Theory”, Springer-Verlag New York, Inc. New York, NY, USA 1995, ISBN: 0-387-94559-8, 1995.

[11] CAO, L. J, TAY, F. E. H “Support Vector Machine With Adaptive Parameters in Financial Time Series Forecasting”, Ieee Transactions On Neural Networks, Vol. 14, No. 6, November 2003, 2003.

[12] Chung, Wai-Chung. Chang, Chuan-Yu. Ko, Chien-Chuan. “A SVM-based Committee Machine for Prediction of Hong Kong” Horse Racing. IEEE, 2017 10th International Conference on Ubi-media Computing and Workshops (Ubi-Media). 2017.

[13] Shin, Jun-Hyun. Lee, Yun-Seong. Kim, Jin-O. “Fault Prediction of Wind Turbine by Using The SVM Method”. IEEE, 2014.

[14] Richard O.Duda, Peter E.Hart and David G. Stork, "Pattern Classification", 2nd Edition, John Wiley & Sons, 2000.

[15] Rawling, John O, et.al. “Applied Regression Analysis:A Research Tool,Second Edition”, Springer, 1998.

[16] S.I.Abba, Sinan Jasim Hadi, Jazuli Abdullahi. “River water modelling prediction using multi-linear regression, artificial neural network, and adaptive neuro-fuzzy inference system techniques”. Procedia Computer Science 120 (2017) 75–82.

[17] Lelitha Vanajakshi, Laurence R. Rilett. “A Comparison Of The Performance Of Artificial. Neural Networks And Support Vector Machines For The Prediction Of Traffic Speed”. 2004 IEEE Intelligent Vehicles Symposium.

[18] Lelitha Vanajakshi, Laurence R. Rilett. “Support Vector Machine Technique for the Short Term Prediction of Travel Time”. 2007 IEEE Intelligent Vehicles Symposium.

Improving Information Service Performance of Family Businesses through IT Governance

Sandy Kosasi
STMIK Pontianak
Pontianak, West Kalimantan, Indonesia
sandykosasi@yahoo.co.id

Vedyanto
Santu Petrus Junior High School
Pontianak, West Kalimantan, Indonesia
vedy91@gmail.com

Velwin Wibowo
STMIK Pontianak
Pontianak, West Kalimantan, Indonesia
velwin.wibowo@gmail.com

***Abstract*—Improving information service performance is central and immensely influences all business activities, run by family enterprises in particular. Contextually, online businesses have not engaged mapping of transparent and accountable information services. This research aimed to obtain more comprehensive results to ease improvement of IT process values in actualizing information service performance. Besides assessing maturity values and managerial implications, the model of IT governance was recommended based on the relationship of processes of Acquire and Implement (AI) Domain, Monitor and Evaluate (ME) Domain, and others through COBIT 4.1 Framework. Results show that averages of AI and ME Domains are respectively 2.788 and 2.417. The lowest scores of these domains are 2.312 (AI4), 2.416 (AI5), 2.356 (ME1), and 2.235 (ME4). A recommended model of IT governance emphasizes the management of changes through horizontal communication and orientation of the seniority level in the family and synchronization and interoperability of improving information service performance for internal and external needs.**

Keywords—Information Service Performance, Family Businesses, IT Governance, AI, ME

I. INTRODUCTION

Roles and functions of IT are on improvement of quicker, more proper, more relevant, more accurate information services of all stakeholders [1]. It pertains more particularly to efficiency, transparency, and acceleration of decision making, empowerment of human resources, and transformation of digital businesses [2]. Information service levels are associated with ownership of proper IT governance based on organizational needs and behavior [3]. IT governance can assure effective, efficient process performance of IT businesses. It also enhances expectation values of all stakeholders [4]. IT governance crucially provides guarantees of achieving business goals and preventing risks of failure or inappropriateness of each business process by noting appropriateness of data management. Moreover, it ascertains that IT empowerment can be reached consistently in synchronization and interoperability through alignment of business strategy and IT when proving information services of all business processes [5]. The other functions of IT governance are specifying decision rights and accountability framework to direct information service attitude based on necessities and determining those contributing systematically to decision making [6]. IT governance is required since IT has become the main stimulus of making business transformation, operational cost decrease, innovation, and efficiency through improvement of information service performance [7]. The fact becomes an important implication for the management to particularly actualize proper IT governance and needs of enterprises operating family businesses in West Kalimantan, Indonesia.

Professionally managing IT governance of family businesses is alike that of other public business organizations [8]. Every family business enterprise also has IT departments managing IT resources when implementing management of application portfolios of information services to support all operational activities, management lines, and top management [9]. Possessing strategic roles actualizes integrated business communication through services and exchange of information more interactively [10]. Effective information service performance is a strategic asset enhancing business performance and users' additional values [11]. It further influences performance of business activities [12].

Preliminary survey results show that the majority of family business enterprises in West Kalimantan have possessed IT departments. Nevertheless, some of such departments are still integrated with others. The use of information services tends to be central and can only be accessed by certain departments. In addition, the management system is oriented toward a seniority factor and ownership of processes, procedures, and mechanisms of departments on the basis of family stratification. Next, decisions of short-term IT investment focusing on operational budget are unable to give overall values. Lastly, there is no information mapping on the relationship between the use of IT and information service performance.

The present condition evidently requires fine information service performance. Availability of good solutions of service system remains an essential need of achieving organizational performance indicators. There should also be standardization of information service system between organizations and online service providers, adequate security, integrity, and availability of information. The fact reflects the two domains such as (a) procurement and implementation of solution software of IT infrastructure system and (b) control and evaluation of information service performance at current and upcoming time.

Several previous studies indicate that existence of IT is getting more strategic in forms of conformity, reliability, and innovation when creating strategic values of information services in online businesses [13]. IT investment management is prone not to focus on IT governance. The lack of knowledge and IT governance results in selection of IT processes not fulfilling the identification capability and implementation to reach conformity of business strategy and IT [14,15]. It is noted that IT governance is limited to fulfillment of internal information services. Therefore, perspectives of strategic needs are ignored [16]. Information service performance immensely determines success of business competitiveness [17]. Measurement of maturity values of IT governance of family businesses in terms of Plan and Organize (PO) and Deliver and Support (DS) Domains is generally at the second level (repeatable but intuitive) despite the fact that the average is at the third level (defined process) [18].

This research was the continuation of the previous study [18] in terms of AI and ME Domains. It aimed to obtain more comprehensive results so that IT processes could be improved to actualize information service performance. Apart from assessing maturity levels, this research intended to cognize managerial implications and recommended a new model of IT governance processes so that the basis of improving information service performance was guaranteed. The research results additionally assured procurement system and implementation of IT governance to ease control and evaluation of information processing system in improving information service performance of family businesses.

II. Literature Review

A. IT Governance

IT governance is the form of responsibility of the top management and integral departments of organizational governance, covers decision rights and accountability framework influencing behavior using IT properly, and ensures effective, efficient IT goals [19]. IT governance reflects responsibility of the board of commissioners, managers, and IT management when formulating organizational structures of business management, business strategy, IT, and processes of IT management to achieve organizational goals, and ensures the implementation [20]. Building organizational relationships through IT governance can only be reached through interrelationships of structures, processes, and mechanisms [21].

B. IT Maturity Levels

Maturity levels of IT governance indicate existence of problems and determine improvement priority. Maturity values are designed as IT process profiles. Accordingly, description of existing and upcoming probability is cognized [22, 23]. Maturity values fulfill maturity criteria of several levels despite similar processes [24]. Through maturity levels of IT governance, current and expected conditions of IT governance are measurable [25].

C. AI Domain

AI Domain includes processes of recognizing feasibility and cost saving solutions technically, guaranteeing development of cost saving applications, and effective, efficient development of infrastructure allowing procurement and maintenance of IT skills, minimizing IT risks, and controlling changes of infrastructure, applications, and technical solutions. Acquiring and implementing strategy should be planned, communicated, and managed. Besides, IT infrastructure should be functioned as it is [26]. AI Domain consists of AI1 (identifying automatic solutions), AI2 (acquiring and maintaining application software), AI3 (acquiring and maintaining IT infrastructure), AI4 (allowing operation and use), AI5 (fulfilling IT resources), AI6 (managing changes), and AI7 (installation, solutions, and changes) [27].

D. ME Domain

ME Domain includes strategy, tactics, and IT mechanism identification contributing based on the conformity control of achievement of business goals. Furthermore, in actualizing strategy, observation and evaluation should be planned, communicated, managed, and prepared. Also, IT infrastructure should be functioned as it is [26]. IT processes of ME Domain comprise ME1 (controlling and evaluating IT performance), ME2 (controlling and evaluating an internal control), ME3 (ensuring fulfillment of external needs), and ME4 (providing IT governance) [27].

III. Research Method

This research applied a quantitative, qualitative method to obtain comprehensive, valid, reliable, objective data [28, 29]. Such the combination method also applied convergent triangulation and an explanatory design through a follow-up explanation model [30]. The first step was to collect data used for computation and quantitative analysis. Identifying parts of quantitative data supported with qualitative data was conducted afterwards. Qualitative data were used to interpret quantitative data through explanation [30]. Obtained results of quantitative data analysis were computed by using Likert Scales consisting of four choices such as Strongly Disagree (SD), Disagree (D), Agree (A), and Strongly Agree (SA) [31]. However, this research used even scales to evade the tendency of respondents to be neutral. Next, obtained results were processed with explanation of informants to strengthen qualitative research results.

Primary data were obtained through questionnaires using the analysis unit of family business enterprises applying online

business marketing. Answers were given by 65 respondents working at 82 family business enterprises. Meanwhile, for triangulation, qualitative data were collected through in-depth interviews and Focus Group Discussions (FGDs). In order to calculate maturity values of IT processes, COBIT 4.1 Framework was in use. This framework can bridge gaps of the two kinds of controls through control objective levels such as activities/tasks, processes, and domains. The main characteristics are on businesses and orientation of business processes based on measurement so that accurate results are achievable [32].

It was started by finding out averages and producing maturity levels of IT processes. Produced results were important for accumulation of data obtained from each process by referring to AI and ME Domains as representation of gap values. Next, aggregation of maturity levels was measured with arithmetic means. The calculation result indicated tendency levels of changes. In order to determine maturity levels of IT processes, index = {∑ (sum of answers x maturity values) : (number of questions x number of respondents)} was the formula in use. Calculation results were mapped by referring to rounding index scales [31, 32]. Results of aggregation values were presented in forms of the table and the radar graphic (see Table I).

TABLE I. ROUNDING INDEX SCALES

Scales	Level of Maturity Model
4.51 – 5.00	5 – Optimized
3.51 – 4.50	4 – Managed and Measurable
2.51 – 3.50	3 – Defined Process
1.51 – 2.50	2 – Repeatable but Intuitive
0.51 – 1.50	1 – Initial/Ad Hoc
0.00 – 0.50	0 – Non-existent

IV. RESULTS AND DISCUSSIONS

Through calculation, the current maturity values of AI and ME Domains are less than 2.50 and do not reach the third level (defined process) yet. Such values are at AI4 (allowing operation and use), AI5 (fulfilling IT resources), ME1 (controlling and evaluating IT performance), and ME4 (providing IT governance). They are respectively 2.312, 2.416, 2.356, and 2.235 (see Tables II and III).

TABLE II. MATURITY LEVEL GAPS OF AI DOMAIN PROCESSES

Domain	Process	Testing Result	Maturity Level
AI1	Identifying automatic solutions	2.956	3
AI2	Obtaining and maintaining application software	2.825	3
AI3	Obtaining and maintaining IT infrastructure	3.325	3
AI4	Allowing operation and use	2.312	3
AI5	Fulfilling IT resources	2.416	3
AI6	Managing changes	2.723	3
AI7	Installation, solutions, and changes	2.960	3

TABLE III. MATURITY LEVEL GAPS OF ME DOMAIN PROCESSES

Domain	Process	Testing Result	Maturity Level
ME1	Controlling and evaluating IT performance	2.356	3
ME2	Controlling and evaluating an internal control	2.565	3
ME3	Ensuring fulfillment of external needs	2.512	3
ME4	Providing IT governance	2.235	3

A computed average of AI Domain is 2.788. However, this score is still low to fulfill a defined process standard. Existing processes have not significantly contributed to stated maturity values yet. For AI Domain, maturity values which are less than 2.51 are at AI4 and AI5 Processes. These two processes give an indication that IT for online businesses of family enterprises does not possess a clearly defined process on operational needs yet to reach a good level of information service performance. Specifically, there are no procedure manuals with clear, structured standardization in managing IT resources. In addition, clear IT description on training materials does not exist. Hence it is hard to understand and apply. Moreover, there are no appropriateness of process standards, procedure manuals, and structured documentation system. Following these, IT trainings tend to vary without comprehensive mechanisms and clear structures. In developing IT, there is no similar approach of all business functions and work units. The orientation is only on each individual. Finally, planning and socialization on trainings, IT implementation, and processes tend to be inconsistently conducted and be dependent on each individual's condition (see Figure 1).

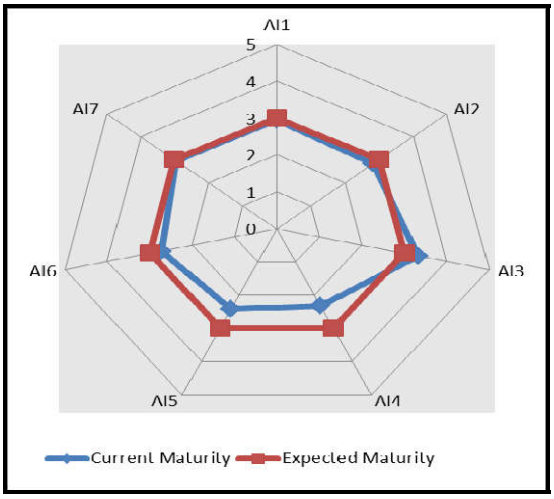


Fig. 1. Maturity Level Model of AI Domain

Meanwhile, another calculation result shows that the average of ME Domain is 2.417. This score is still less than the one in the range of 2.51. In other words, it is at a repeatable but intuitive level. Despite the fact that there are two processes (ME2 and ME3) reaching the score that is greater than 2.51, they have not significantly given contributions on influences of other processes. This condition represents that the use of IT is more centralized to needs of individuals and work units. All IT

processes are at a repeatable but intuitive position. Also, all family business enterprises operating online businesses do not have clear, effective planning on an internal control program through well-defined observation yet. Besides, effective IT governance defining organizational structures, processes, leadership, roles, and responsibility, and aiming to ascertain that IT investment has conformity with strategy and goals is absent. Based on observation, IT process controls are not concerned well. Ultimately, possession of an internal control ensuring operational effectiveness based on valid rules does not exist (see Figure 2).

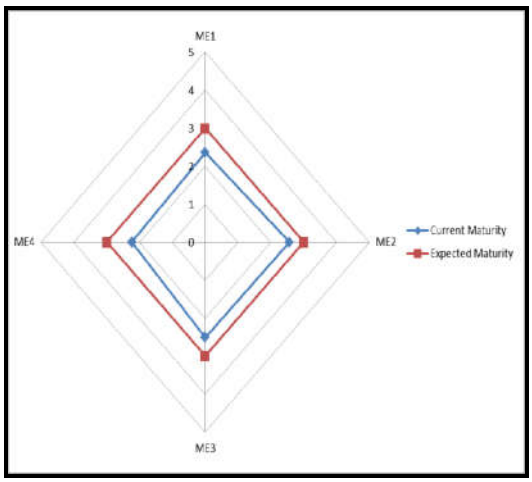


Fig. 2. Maturity Level Model of ME Domain

Research results indicate that based on the description, IT processes are dissimilar and do not reach a defined process level yet. This becomes an essential implication for managers of family business enterprises operating online businesses that information service performance should be concerned seriously.

A way to improve AI4 Domain (allowing operation and use) is planning of addition, implementation, and maintenance of IT infrastructure. This can fulfill business functions and technical needs based on IT directions. Implementing an internal control, security, and auditing size while configuring, integrating, and maintaining infrastructure of hardware and software to protect resources is another requirement. Assuring availability and integrity is also certainly needed. Developing the strategy of planning of the infrastructure treatment and ensuring that control changes are in line with management procedures of organizational changes should be further conducted. Next, regarding AI5 Domain (fulfilling IT resources), there should be allocation of procedure controls of IT processes. Procurement comprises computing resources, servers, network devices, operation system, application software, programs, software modules, internet, consultants, and other external human resources. Management and procurement of IT can be conducted by making application portfolios independently, renting the internet connection, holding IT workshops, cooperating with the third parties and IT industries, and exchanging services with vendors.

On the other hand, improvement of ME1 Domain (controlling and evaluating IT performance) is through redefinition of indicators of information service performance including arrangement of schedules, techniques and mechanisms of measurement, reports of analysis results, and analysis of all measurement results. Mechanisms of controls and evaluation are required to ascertain IT implementation based on directions and policy of the management of online businesses. Meanwhile, ME4 Domain (providing IT governance) can be improved through IT processing. The management should take remedial actions on inappropriate processes through existing standards to prevent similar cases in the future. Improving effectiveness of IT governance is through definition of organizational structures, IT processes, leadership, roles, and responsibility assuring that IT investment and capabilities to determine measurement indicators of information service performance are in line with business strategy and IT.

Information service performance should be completed with readiness to implement IT governance by referring to obtained, expected maturity values. Updates of investigating all IT processes are required to improve information service performance. It is noted that in operating online businesses, availability and easiness to get information of products should exist. Moreover, improving IT governance can ensure control mechanisms and evaluation of IT investment producing business values and minimizing IT risks.

Thus, initially, improving maturity values of IT governance is necessary. Maturity of AI and ME Domains is related to the one of PO and DS Domains. All of these domains influence each other through IT processes. Interrelated processes aim to improve information service performance through alignment of business strategy and IT of family enterprises operating online businesses.

A recommended model of IT governance is on domains with IT processes possessing values which are less than 2.51, i.e. AI4, AI5, ME1, and ME4. These four processes are very important factors since they have direct relationships with PO and DS Domains. However, they need improvement. This activity requires commitment, readiness, and assurance of implementing change management oriented toward effectiveness of IT governance.

The following description of a recommended model of IT governance is used to improve expected maturity values. Each IT domain represents a number of detailed activities on objective controls of inputs and outputs. Maturity of IT governance is the representation and guarantee of improving information service performance.

1) AI4 Domain (allowing operation and use) is related to other objective controls. The input objective control consists of PO10 (detailed project plans), AI1 (business requirements), AI2 (application software knowledge), AI3 (infrastructure knowledge), AI7 (known and accepted errors), and DS7 (ducumentation updates). Conversely, the output objective control comprises AI7, DS4, DS8, DS9, DS11, DS13 (users, operational supports, and technical and administration manuals), DS7 (knowledge transfer requirements), and DS7 (training materials).

2) AI5 Domain (fulfilling IT resources) is related to other objective controls. The input objective control consists of PO1 (IT acquisition strategy), PO8 (acquisition standards), PO10 (detailed project plans), AI1 (feasibility studies), AI2-3 (procurement decisions), and DS2 (supplier catalogs). In contrast, the output objective control covers DS2 (third party relationships), AI7 (procured items), and DS2 (contractual arrangements).
3) ME1 Domain (controlling and evaluating IT performance) is related to other objective controls. The input objective control comprises PO5 (cost-benefit reports), PO10 (project performance reports), AI6 (changing status reports), DS1-13 (process performance reports), DS3 (performance and capacity plans), DS8 (user satisfaction reports), ME2 (reports on effectiveness of IT controls), ME3 (reports on compliance of IT activities), and ME4 (reports on IT governance status). On the other hand, the output objective control consists of PO1, PO2, DS1 (performance inputs to IT planning), PO4, PO8 (remedial action plans), PO9 (historical risk trends and events), and ME2 (process performance reports).
4) ME4 Domain (providing IT governance) is related to objective controls. The input objective control covers PO4 (IT process framework), PO5 (cost-benefit reports), PO9 (risk assessment and reporting), ME2 (reports on effectiveness of IT controls), and ME3 (catalogs of legal and regulatory requirements related to IT service deliveries). Meanwhile, the output objective control consists of PO4 (process framework improvement), PO1, ME1 (reports on IT governance status), PO5 (IT-enabled business investments), PO1 (strategic enterprise directions for IT), and PO9 (enterprise appetites for IT risks).

V. Conclusion and Future Research

Improving information service performance of family enterprises operating online businesses in West Kalimantan should be through effective IT governance referring to relationship conformity of business strategy and IT. Conducting change management having horizontal communication and not focusing on seniority of family should be concerned. These are crucial activities of making strategy of implementing IT governance through integration of IT processes. Integration mechanisms of processes of businesses and IT are important parts of achieving synchronization and interoperability when improving information service performance for internal and external needs.

Maturity values of IT governance of AI and ME Domains show these averages: 2.788 and 2.417. These two values do not reach a defined process position or an expected maturity level yet. In fact, the one of AI Domain does not reach 3.00 or even exceeds 3.51. Meanwhile, the maturity value of ME Domain is still lower and is at a repeatable but intuitive position. It is recommended that AI4 (allowing operation and use), AI5 (fulfilling IT resources), ME1 (controlling and evaluating IT performance), ME4 (providing IT governance), and other IT processes should be interrelated. Improvement of maturity values should be immediately conducted to achieve consistency values of processes.

This research can be continued with analysis of IT governance on interrelationships of PO, DS, AI, and ME Domains to obtain Key Performance Indicator (KPI) and Key Goal Indicator (KGI) of each IT process. Determining KPI and KGI through combination of gaps of PO, DS, AI, and ME Domains can provide more comprehensive information on weakness aspects of IT processes of IT governance of family enterprises pertaining to online business operation in West Kalimantan. Hence, this stimulus is important for achievement of conformity of business strategy and IT when supporting improvement of information service performance.

References

[1] S. Héroux, and A. Fortin, "The Influence of IT Governance, IT Competence and IT-Business Alignment on Innovation," Cahier de recherche, 4, 2016, pp.1-36.

[2] T. Pham, D.K. Pham, and A.T. Pham, "From Business Strategy to Information Technology Roadmap: A Practical Guide for Executives and Board Members," CRC Press Taylor & Francis Group, 2013.

[3] V. Raodeo, "IT Strategy and Governance: Frameworks and Best Practice," International Journal of Research in Economics & Social Sciences, 2(3), 2012, pp.49-59.

[4] R.S. Debreceny, and G.L. Gray, "IT Governance and Process Maturity: A Multinational Field Study," Journal of Information Systems, 27(1), 2013, pp.157-188.

[5] B.S. Sidhu, and K. Gupta, "A Critical Study of IT Transformation Practices to Achieve Business IT Alignment," International Journal of Emerging Trends in Science and Technology, 2(4), 2015, pp.2289-2299.

[6] W. Ly Teo, A. Manaf, and P. Lai Fong Choong, "Perceived Effectiveness of Information Technology Governance Initiatives among IT Practitioners," International Journal of Engineering Business Management, 15(19), 2013, pp.1-9.

[7] R. Pereira, and M.M. Silva, "A Literature Review: Guidelines and Contingency Factors for IT Governance," European, Mediterranean & Middle Eastern Conference on Information Systems, 2012, pp.342-360.

[8] I. Robinson, and M. Jodha, "Building and Maintaining Effective Mechanisms for Implementing IT Governance," ISACA Journal, 1, pp.2013, 1-3.

[9] J. Iden, and T.R. Eikebrokk, "Using the ITIL Process Reference Model for Realizing IT Governance: An Empirical Investigation," Information Systems Management, 31(1), 2014, pp.37-58.

[10] A. López-Poveda, L. Rusu, and P. Johannesson, "A Maturity Model for Analyzing Strategic IT Management from a Service Perspective," AMCIS 2011 Proceedings - All Submissions, 2011, pp.1-11.

[11] M. Marrone, and L.M. Kolbe, "Einfluss Von IT-Service-Management-Frameworks Auf die IT-Organisation," Wirtschaftsinformatik, 53(1), 2011, pp.5-19.

[12] A.A. Neff, F. Hamel, T.P. Herz, F. Uebernickel, and W. Brenner, "IT Governance in Multi-Business Organizations: Performance Impacts and Levers from Processes, Structures, and Relational Mechanisms," 46th Hawaii International Conference on System Sciences, IEEE Computer Society, 2013, pp.4466-4475.

[13] B. Bigliardi, L. Cassia, A. De Massis, and F. Frattini, "Technology Strategy in Family Business: A New Avenue for Research," International Journal Entrepreneurship and Innovation Management, 17(1), 2013, pp.1-7.

[14] L.P. Steier, J.J. Chrisman, and J.H. Chua, "Governance Challenges in Family Businesses and Business Families,", Entrepreneurship Theory and Practice, 2015, pp.1265-1280.

[15] S. Horan, G. Szundi, C. Edward Van Deman, and K.P. Whalen, "Technology and the Family Office," The Journal of Wealth Management, 16(2), Banking Information Source, 2013, pp.99-111.

[16] H. Rui, Z. Robert W, P. R. Leon, "Influencing the Effectiveness of IT Governance Practices through Steering Committees and Communication Policies," European Journal of Information Systems, 19, 2010, pp. 288-302.

[17] J. Sandlin, "Family Business Adaptation to Disruptive Technology: Case Studies on Family Businesses in Handling the Challenges of Disruptive Technology and Maintaining Competitive Advantage within a Swedish Market," Jonkoping University, 2017.

[18] S. Kosasi, and H. Prabowo, "Evaluation of Information Technology Governance Implementation in Business Enterprises: Case: Family Business Enterprises Pontianak, West Kalimantan, Indonesia," International Conference on Information Management and Technology, 2016, pp.1-6.

[19] S. De Haes, W.V. Grembergen, and R.S. Debreceny, "COBIT 5 and Enterprise Governance of Information Technology: Building Blocks and Research Opportunities," Journal of Information Systems, 27(1), 2013, pp.307-324.

[20] L. Janahi, M. Griffiths, and H. Al-Ammal, "A Conceptual Model for IT Governance in Public Sectors," Fourth International Conference on Future Generation Communication Technology, 2015, pp.1-9.

[21] M.A.H. Altemimi, and M.S. Zakaria, "Developing Factors for Effective IT Governance Mechanism," 9th Malaysian Software Engineering Conference (MySEC), 2015, pp.245-251.

[22] A. Joshi, L. Bollen, H. Hassink, S. De Haes, and W.V. Grembergen, " Explaining IT Governance Disclosure through the Constructs of IT Governance Maturity and IT Strategic Role," Information & Management, 2017, pp.1-61.

[23] M.A. Al-sa'eed, S.M. Al-mahamid, and R.M.H. Al-sayyed, "The Impact of Control Objectives of Information and Related Technology (COBIT) Domain on Information Criteria and Information Technology Resources," Journal of Theoretical and Applied Information Technology, 45(1), 2012, pp.9-18.

[24] G.J. Selig, "IT Governance-An Integrated Framework and Roadmap: How to Plan, Deploy and Sustain for Improved Effectiveness," Journal of International Technology and Information Management, 25(1), 2016, pp.55-76.

[25] D. Roger S, and G. Glen L, "IT Governance and Process Maturity: A Multinational Field Study," Journal of Information Systems, 27(1), 2013. pp.157-188.

[26] N. Rezaei, "The Evaluation of Implementing IT Governance Controls", Journal of Applied Business and Finance Researches, 2(3), 2013, pp.82-89.

[27] IT Governance Institute, "COBIT 4.1: Framework, Objektive Controls, Management Guidelines, Maturity Models", ITGI, 2007.

[28] J.W. Creswell, "Research Design: Qualitative, Quantitative, and Mixed Methods Approaches," Fourth Edition, California: SAGE Publications, Inc., 2014.

[29] S.N. Hesse-Biber, and B. Johnson, "The Oxford Handbook of Multimethod and Mixed Methods Research Inquiry," Oxford Library of Psychology, 2015.

[30] B. Agustinus, "Penelitian Kualitatif: Metodologi, Desain, dan Teknik Analisis Data dengan NVIVO 11 Plus," Mitra Wacana Media, 2016.

[31] U. Sekaran, & R. Bougie, "Research Methods for Business: A Skill Building Approach (Sixth Edition ed.)," United Kingdom: John & Wiley & Sons, Ltd., 2013.

[32] G. Mangalaraj, A. Singh, and A. Taneja, "IT Governance Frameworks and COBIT-a Literature Review," Twentieth Americas Conference on Information Systems, Savannah, 2014, pp. 1-10.

[33] R.A. Khther, and M. Othman, "COBIT Framework as a Guideline of Effective IT Governance in Higher Education: A Review," International Journal of Information Technology Convergence and Services 3(1), 2013, pp.21-29.

[34] J. Orozco, A. Tarhini, and T. Tarhini, "A Framework of IS/Business Alignment Management Practices to Improve the Design of IT Governance Architectures," International Journal of Business and Management, 10(4), 2015, pp.1-12.

Measuring Maturity Index of Risk Management for IT-Governance Using Fuzzy Ahp and Fuzzy Topsis

Reza Amalia Priyantina
Department of Technology Information Management
Institut Teknologi Sepuluh Nopember (ITS)
Surabaya, Indonesia
reza.17092@mhs.its.ac.id

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember (ITS)
Surabaya, Indonesia
riyanarto@if.its.ac.id

Abstract - Increased business competition among the companies demand them to develop more. The development is not separated from the business risk and weaknesses. The business risks and weaknesses can be known and evaluated quickly using SOP (Standard Operating Procedure) which is simulated through YAWL application. Once the simulated weaknesses are identified, then it is conducted the index maturity of risk management to determine the company's understanding and concern on the existence of risks in the company's business process and risk management responsibilities within the organization itself. The method of measuring the evaluation of maturity index on risk management mostly uses measurement step by assessing and weighing the result from observation, interview, survey, etc., but if the data is not quite good then it will cause problem. Therefore, we need a new model for measuring and knowing the risk alternatives faster to be able to initiate risk-prevention. In this research, fuzzy logic model is used in making decision of the biggest risk alternative and to be handled for the first time. Fuzzy AHP will be used to calculate the value of the criteria, taken from the maturity index COBIT and Fuzzy TOPSIS to determine the selected alternative rating. It is observed that Managed and Measurable maturity index have the smallest weight of 0.0355 compared to other criteria (That means the company does not reach the level of maturity to understand risks in managed and measure) and Database Server achieves very serious attention because the score is very high compared to the other alternatives of 1.1186.

Keywords: Risk Management, Maturity Index, IT Governance, COBIT, Fuzzy AHP, Fuzzy TOPSIS.

I. INTRODUCTION

The competitive and rapidly changing business environment demands the company develop more. The development is conducted by increasing the productivity, improving the quality of business process and reducing the cost of business work. In addition, reducing company's weaknesses and business risks is also important. This step can be conducted well if the company is able to evaluate quickly the whole running process. Evaluation of running business processes can be performed by implementing Business Process Management (MPB). Business Process Management is conducted with the aim of obtaining the efficient business process for the organization.

The business process consists of a set of activities performed in the company. Usually, business process in the company is referred to as SOP (Standard Operating Procedure). This SOP will be simulated and evaluated through YAWL application software to know the running business process activity and the record will generate event log. Event log records all sequences of activity that runs from the initial to the final activity. It is made a simulation to predict the weakness of company's SOP that have the potential as a source of risk from the event log.

Risk-based IT management is very important because the successful companies are able to understand and manage the risks that may occur and threaten the information technology asset in the company.

Information technology governance has a wide scope and concentrates on the performance and transformation of information technology to achieve the company's business objectives, both internal and external business [1]. There are five 5 focus areas of IT Governance those are: Strategic alignment, Value delivery, Resource management, Risk management, and performance measurement.

Risk management is needed to be conducted to reduce losses and increase opportunities [2]. In this study, the used risk management supporter process (based on COBIT) are: AI2 (obtaining and maintaining application software), DS5 (ensuring system security), DS11 (managing data) and DS12 (managing the physical environment).

Maturity index is needed to be measured to make the management understands the level of company maturity position at the moment, and continuously tries to improve it so that the aspect of governance on company's information technology can run effectively [3].

Method of maturity index evaluation measurement on risk management has used many measurement steps by assessing and weighting the result of observation, interview, survey, etc., but if we obtain not adequately good data it will cause some problems. The existing data is sometime inadequate for real-life problem solving since human calculation on preferences is unclear and unpredictable with the exact numeric score [4] and [5]. So, it needs new a model to perform measurement. This research uses fuzzy logic model in the decision making of structured choice. Fuzzy logic helps measure the concept of uncertainty related to subjective human. Fuzzy AHP is used to calculate the value of the criteria and Fuzzy TOPSIS is used to decide the ranking of selected alternatives.

II. BASIC THEORY

A. Business Process Management

Business Process Management is method and technique used to support business process in terms of modeling, controlling, and analyzing process that involve process subject, organization, application, document and other information input [6]. Business process management aims to get a good and efficient business process for the organization.

B. Risk Management

One of the main focuses of IT governance is risk management, it is very important and interesting topic for many organizations as it covers the organization's success in reaching its early stated goals and eliminating possible failures [7]. So, risk management has an important role in the overall management system [8].

C. Maturity Index

Maturity index is an index of maturity level in COBIT. Maturity model for management and control over IT process is based on organizational evaluation method, so it can be judged from the level of maturity that does not exist (0) to be optimized (5). [9]

Control Objectives for Information and related technology (COBIT) is created and distributed by Information System Audit and Control Association (ICASA) [10] COBIT is a standard in the framework domain consists of a set of IT processes that present controllable and structured activities [11]. COBIT framework provides domain process models that are commonly discovered in IT Activities, such as: Plan and Organize (PO), Acquire and Implement (AI), Deliver and Support (DS) and Monitor and Evaluate (ME).

IT governance focus areas on Risk Management has supporting processes as follows:

TABLE I. IT TABLE PROCESS SUPPORTS IT GOVERNANCE BASED ON COBIT

(Source: [11])

Governance Focus Area	Supporting processes	
	Primary	Secondary
Risk Management	PO4, PO6, PO9, The DS2, DS4, DS5, DS11 DS12, ME2 ME3 ME4,,	PO1, PO2, PO3, PO7 PO8,, AI1, AI2, PO10, AI4, AI7, DS3, DS7, DS9, DS10, ME1

D. Fuzzy Analitical Hierarchy Process (FAHP)

Fuzzy Analitical Hierarchy Process (FAHP) Method is proposed for the first time by Chang and this is a development of AHP method created by Saaty which FAHP is a method consists of AHP matrixes represented or filled by fuzzy number [12][14].

FAHP uses fuzzy ratio called by Triangular Fuzzy Number (TFN) and used in the calculation process. Triangular Fuzzy Number consists of three function criteria scores, namely the lowest score (l), the middle score (m), and the highest score (u).

The step in performing FAHP is as follows [4]:

1. Arranging the problems in the form of hierarchy

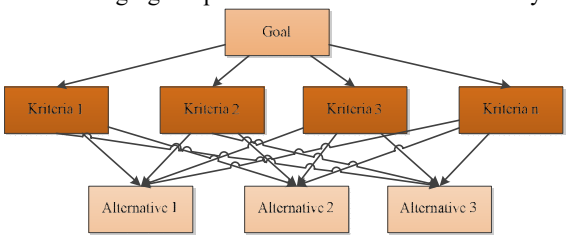


Fig. 1. Problem Hierarchy Structure

2. Establishing the weight of each predefined criteria using fuzzy triangle (TFN) score that indicates the comparing of importance among the criteria.

TABLE II. VARIABLE LINGUISTIC SCORE WITH TRIANGULAR FUZZY NUMBER [13]

Variable Linguistic	Value TF	Value Invers TF
Equal Importance (EI)	1 1 1	1 1 1
Moderate Importance (MI)	1 3 5	1/5 1/3 1
Strong Importance (SI)	3 5 7	1/7 1/5 1/3
Very Strong Importance (VSI)	5 7 9	1/9 1/7 1/5
Extreme Importance (EI)	8 9 10	1/10 1/9 1/8

Triangular Fuzzy Figure in table 1 is denoted by M = {l, m, u}, which M is a matrix of fuzzy number consists of l (lowest score), m (medium/nearest score) and u (biggest score).

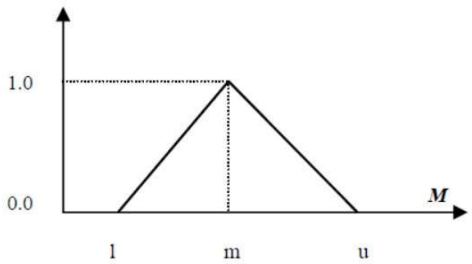


Fig. 2. Triangular Fuzzy Number (TFN)

If X={x1,x2,x3,..xn} indicates a set of objects, and Y={y1,y2,y3,…yn} indicates stipulated goals. If there are number of analyzed m criteria, then it is obtained that M yi1, M yi2, M yi.3 ... M yi.mi = 1,2, ...,n, which Myi j (j = 1, 2, ..., n) triangular fuzzy number.

3. Determine fuzzy synthesis score with the criteria of I using the formula of 1.

$$S_i = \sum_{j=1}^{m} M_{gi}^{j} \times \left[\sum_{i=1}^{n} \sum_{j-1}^{m} M_{gi}^{j} \right]^{-1}$$

Formula 1

To obtain $\sum_{j=1}^{m} M_{gi}^{j}$, it is calculated using formula 2.

$$\sum_{j=1}^{m} M_{gi}^{j} = \left(\sum_{j=1}^{m} l_j, \sum_{j=1}^{m} m_j, \sum_{j=1}^{m} u_j \right)$$

Formula 2

To obtain $\left[\sum_{j=1}^{n} \sum_{j=1}^{m} M_{gi}^{j} \right]^{-1}$, it is conducted sum of fuzzy score using formula 3.

$$\sum_{j=1}^{n} \sum_{j=1}^{m} M_{gi}^{j} = \left(\sum_{j=1}^{n} l_j, \sum_{j=1}^{n} m_j, \sum_{j=1}^{n} u_j \right)$$

Formula 3

Next, count the inverse score of matriks in formula 3 using formula 4.

$$\left[\sum_{j=1}^{n} \sum_{j=1}^{m} M_{gi}^{j} \right]^{-1} = \left(\frac{1}{\sum_{j=1}^{n} l_j}, \frac{1}{\sum_{j=1}^{n} m_j}, \frac{1}{\sum_{j=1}^{n} u_j} \right)$$

Formula 4

4. Calculate level of possibility, which A1 = (l1, m1, u1) and A2 = (l2, m2, u2) is two of TFN, so level of possibility of A2 = (l2, m2, u2) ≥ (l1, m1, u1) it is from the formulation of formula (5) and (6).

$$V = A_2 \geq A_1 = hgt (A_2 \cap A_1) = \mu_{M2}$$

Formula 5

$$= \begin{cases} 0 & , if \quad m_2 \geq m_1 \\ 1 & , if \quad l_1 \geq u_2 \\ \dfrac{l_1 - u_2}{(m_2 - u_2) - (m_1 - l_1)} & , otherwise \end{cases}$$

Formula 6

5. Compare level of possibility between fuzzy A1 number = (i=1, 2,…k) with the formula 7.

$$P(A \geq A_1, A_2, \ldots A_k) = P[(A \geq A_1)] \text{ and } (A \geq A_2) \text{ and } \ldots \text{ and } (A \geq A_k)] = \min P(A \geq A_1), i = 1,2,\ldots,k$$

Formula 7

With the estimation of d (Ni) = minP (Si ≥ Sj) for j = 1,2, ... j; j ≠ i. so we obtain value of weight vector on the formula (8).

$$W' = (d'(N_1), d'(N_2), \ldots, d'(N_n))^T$$

Formula 8

After the formula (8) is normalized, the obtained normal weight vector can be calculated using formula (9).

$$W = (d(N_1), d(N_2), \ldots, d(N_n))^T$$

Formula 9

Which found W is not a Fuzzy score.

E. Fuzzy TOPSIS

After the criteria calculated using Fuzzy AHP, then it conducted an alternative ranking using TOPSIS method that has been using fuzzy numbers. Fuzzy TOPSIS uses fuzzy triangle number as in illustration 3 to describe the score for each of selected alternative.

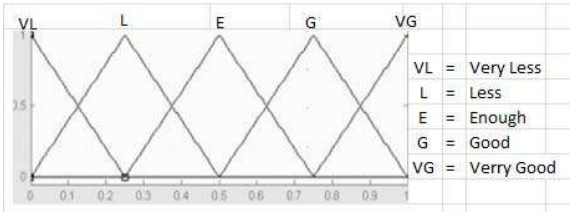


Fig. 3. Fuzzy Number for Alternative Assessment

After each specified alternative is assigned a value, we can calculate the total integral weight for each alternative using the formula (10).

$$x = I(F) = ½ (ac+b+(1-a)a)$$

Formula 10

With h α is a level of optimism by the grade between 0 and 1. After obtaining the total integral weight for each criterion, it is subsequently normalized by the formula (11).

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}}$$

Formula 11

Followed by performing calculation to obtain a normalized weight matrix through formula (12).

$$v_{ij} = W_{ij} * r_{ij}$$

Formula 12

Continued by calculate the score of positive ideal solution and negative ideal solution using formula (13).

$$A^{+} = \{v_1^{+}, v_2^{+}, ..., v_n^{+}\}$$
$$A^{-} = \{v_1^{-}, v_2^{-}, ..., v_n^{-}\}$$

Formula 13

Range between alternatives with positive ideal solution is calculated using the formula (14).

$$D_i^{+} = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^{+})^2}$$

Formula 14

Meanwhile, the alternative range to the ideal solution is calculated by the formula (15).

$$D_i^{-} = \sqrt{\sum_{j=1}^{n} (v_{ij} - v_j^{-})^2}$$

Formula 15

By compare the range with the positive and negative ideal solution, then it is calculated the preference for each alternative by the formula (16).

$$CC_i = \frac{D_i^{-}}{D_i^{-} + D_i^{+}}$$

Formula 16

The calculated value of formula (16) is the final score used to decide the alternative based on the largest value of CC_i.

Combination between the two Fuzzy methods (Fuzzy AHP and Fuzzy Topsis) at this paper, illustrates how fuzzy AHP and Fuzzy Topsis can be integrated to allow a more consistent evaluation and prioritize the smallest weight of the criteria to be improved (with fuzzy AHP) and to determine the ranking of the alternatives that have been selected (with fuzzy topsis).

III. RESEARCH METHODOLOGY

The designed research model is shown in Figure 4.

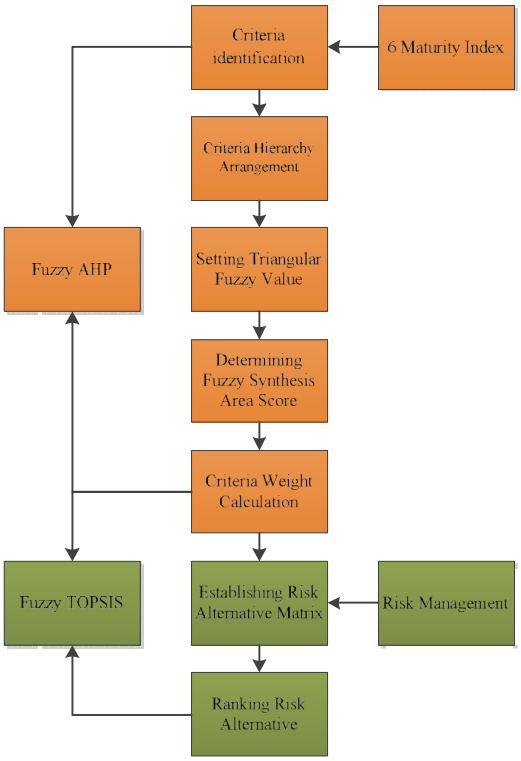


Fig. 4. Research Methode

First, this study was to decide the criteria. The used criteria were taken from 6 COBIT maturity indexes, 0 Non-Existent, 1 Initial/Ad Hoc, 2 Repeatable but Intuitive, 3 Defined, 4 Managed and Measurable 5 Optimized. After obtaining the criteria, the calculation was conducted using FAHP started by making pairwise comparison matrixes from the maturity index to measure the fuzzy synthetic score so we obtain the weight of each criteria.

The calculated weight using FAHP was used to rank the risk management search by using Fuzzy TOPSIS. Risk management in this study included: Hardware (Damaged or Missing), Database Server (vital asset because it was used as a whole sensitive data storage from simple to complex data), SI Application (The main system used by companies to enter data and producing report), Infrastructure, Network and Employees (employees as a whole were trained in the use of Application and security awareness of application usage, data theft, etc.) determined by checking business process using YAWL.

IV. RESULT

The first step is to know the criteria by considering the index value of maturity index in 6 attributes of COBIT maturity includes: 0 Non-Existent, 1 Initial / Ad Hoc, 2 Repeatable but Intuitive, 3 Defined, 4 Managed and Measurable 5 Optimized.

For each used criterion used, the weight of M is given to represent the ratio among the criteria to each other. The ratio between all criteria or attributes is set to get the matrix as in Table 2.

TABLE III. COMPARISON MATRIX OF PAIR FUZZY SCORE OF COBIT MATURITY INDEX.

Criteria	Non-existent			Initial/Ad ho			table but Ins			Defined			and M			Optimised		
Non-existent	1	1	1	1	3	5	3	5	7	1	3	5	1	1	1	1	1	1
Initial/Ad hoc	1/5	1/3	1	1	1	1	1/5	1.3	1	3	5	7	3	5	7	3	5	7
Repeatable but Instuitive	1/7	1/5	1/3	1	3	5	1	1	1	1	3	5	1	3	5	3	5	7
Defined	1/5	1/3	1	1/7	1/5	1.3	1/5	1.3	1	1	1	1	1	3	5	1/5	1/3	1
Managed and Measurable	1	1	1	1/7	1/5	1.3	1/5	1.3	1	1/5	1/3	1	1	1	1	1	1	1
Optimised	1	1	1	1/7	1/5	1.3	1/7	1/5	1/3	1	3	5	1	1	1	1	1	1

Next, it is calculated the weight of fuzzy Synthetic extents (Si) using the formula (1), and the result is seen in table 4.

TABLE IV. SYNTHETIC EXTENTS SCORE

Criteria	L	M	U
Non-existent	8.0000	14.0000	20.0000
Initial/Ad hoc	10.4000	16.6667	24.0000
Repeatable but Instuitive	7.1429	15.2000	23.3333
Defined	2.7429	5.2000	9.3333
Managed and Measurable	3.5429	3.8667	5.3333
Optimised	4.2857	6.4000	8.6667
Number of Columns	36.1143	61.3333	90.6667

Then, it is used the formula (3) to calculate the weight of each criteria level. The generated matrix calculation score from table 4 are calculated by the formula (8) as in table 5. (W ').

TABLE V. WEIGHT VECTOR SCORE

Criteria	on-existen	itial/Ad h	ble but In	Defined	and M	Optimised
Non-existent	1	1.0816	1.0363	0.5426	0.2646	0.5505
Initial/Ad hoc	0.9099	1	0.9569	0.4346	0.1364	0.4280
Repeatable but Instuitive	0.8824	1.0549	1	0.5242	0.2716	0.5291
Defined	1.6690	1.9181	1.7970	1	0.8438	1.1029
Managed and Measurable	2.2911	2.5882	2.3929	1.5763	1	1.2588
Optimised	1.8375	2.1092	1.9655	1.1105	0.6291	1

Then, it is conducted normalization for each criterion, so there is a vector score shown in Table 5.

TABLE VI. WEIGHT NORMALIZATION

Criteria	on-existen	itial/Ad h	ble but In	Defined	and M	Optimised
W'	0.8824	1	0.9569	0.4346	0.1364	0.4280
W	0.2299	0.2605	0.2493	0.1132	0.0355	0.1115

The weight of vector score generated from the calculation using Fuzzy AHP method is used to calculate the alternative ranking value taken from each risk using TOPSIS method. The identification of used alternatives in this study refers to some of the identified risks SOP using YAWL application, such as: Hardware (Damaged or Missing), Database Server (vital asset because it is used as a sensitive data storage from simple data to complex), Information System Application, Infrastructure, Network and Employees (employees as a whole are trained in application use and security awareness of application use, data theft, etc.)

Each risk criteria are assigned using a fuzzy triangle (TFN), as in Table 7.

TABLE VII. ALTERNATIVE VALUE NORMALIZATION

Criteria	Non-existent			Initial/Ad hoc			peatable but Instuiti			Defined			naged and Measura			Optimised		
Infrastructure	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75	0.5	0.75	1	0.5	0.75	1	0.75	1	1
Network	0.25	0.5	0.75	0.5	0.75	1	0.5	0.75	1	0.75	1	1	0.75	1	1	0.75	1	1
Q-Pro Application	0	0.25	0.5	0.5	0.75	1	0.25	0.5	0.75	0.5	0.75	1	0.5	0.75	1	0.75	1	1
Hardware	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75	0.5	0.75	1
Database Server	0	0.25	0.5	0.5	0.75	1	0.25	0.5	0.75	0.25	0.5	0.75	0.25	0.5	0.75	0.5	0.75	1
Employee	0	0.25	0.5	0.25	0.5	0.75	0.25	0.5	0.75	0	0.25	0.5	0	0.25	0.5	0.25	0.5	0.75

Thereafter, the score obtained from Table 7 are calculated using the formula 10 to obtain the integral total value and the obtained X matrix as shown in Table 8.

TABLE VIII. INTEGRAL SCORE TABLE

Criteria	Non-existent	Initial/Ad hoc	peatable but Instuiti	Defined	naged and Measura	Optimised
Infrastructure	0.4375	0.4375	0.4375	0.75	0.75	0.96875
Network	0.4375	0.75	0.75	0.96875	0.96875	0.96875
Q-Pro Application	0.125	0.75	0.4375	0.75	0.75	0.96875
Hardware	0.4375	0.4375	0.4375	0.4375	0.4375	0.75
Database Server	0.125	0.75	0.4375	0.4375	0.4375	0.75
Employee	0.125	0.4375	0.4375	0.125	0.125	0.4375

Normal v weight matrix is calculated using formula (11) and (12) in table 8, so that the result of matrix V can be seen in Table 9 below:

TABLE IX. NORMALIZATION WEIGHT MATRIX

Criteria	Non-existent	Initial/Ad ho	atable but Instu	Defined	ged and Measu	Optimised
Infrastructure	0.1276	0.0758	0.0885	0.0541	0.0170	0.0531
Network	0.1276	0.1299	0.1517	0.0699	0.0219	0.0531
Q-Pro Application	0.0365	0.1299	0.0885	0.0541	0.0170	0.0531
Hardware	0.1276	0.0758	0.0885	0.0316	0.0099	0.0411
Database Server	0.0365	0.1299	0.0885	0.0316	0.0099	0.0411
Employee	0.0365	0.0758	0.0885	0.0090	0.0028	0.0240

Then, it is calculated the value of ideal positive solution score and ideal negative score in the normalized weight matrix in table 8 using the formula (13), the distance between the positive ideal solution alternative (Di^{+}) is calculated using the formula (14), while the range alternative of negative ideal solution (Di^{-}) is calculated by using the formula (15).

By comparing the range between a positive ideal solution and a negative ideal solution, a preference score for each of alternatives determined in Table 10 is calculated using formula (16).

TABLE X. RISK ALTERNATIVE RANKING

Criteria	Cci	Rank
Infrastruktur	1.0848	4
Jaringan	1.0000	5
Aplikasi SI	1.1122	2
Hardware	1.0932	3
Database Server	1.1186	1
Karyawan	0.1420	6

V. CONCLUSION

This research is to determine and evaluate maturity index and risk management in the implementation of IT governance using Fuzzy AHP and Fuzzy TOPSIS at PT. X. The results indicate that Managed and Measurable maturity index has the smallest weight of 0.0355 compared to other criterian (That means companies do not reach maturity levels to understand risks in "managed and measure: Management monitors and measures compliance to procedures and takes action if the process cannot be done effectively") and Database Server obtains very serious attention because the score is very high compared to another alternative with 1.1186. Then, it is necessary to take strategic steps from the company to prevent or improve the risk that may occur.

REFERENCES

[1] DS. Kerr and US. Murthy, "The importance of the CobiT zframework IT processes for effective internal control over financial reporting in organizations: An international survey," *Inf Manag*, 2013, pp.590-597.

[2] MR. Fazlida and J. Said, "Information Security: Risk, Governance and Implementation Setback," Procedia Econ Financ, April 2015, pp.243-248.

[3] Johan J. C. Tambotoh and R. Latuperissa, "The Application for Measuring the Maturity Level of Information Technology Governance on Indonesian Government Agencies Using COBIT 4.1 Framework," Sci Res Intell Inf Manag, vol. 6, pp. 12-19, 2014.

[4] U. Yudatama and R. Sarno, "Evaluation Maturity Index and Risk Management for IT Governance Using Fuzzy AHP and Fuzzy Topsis," Int Semin Intell Technol Its Appl, August 2015, pp. 323-328.

[5] U. Yudatama and R. Sarno, "Priority Determination for Higher Education Strategic Planning Using Balanced Scorecard, FAHP and TOPSIS (Case study: XYZ University)," IOP Conf Ser Mater Sci Eng, 2016;105(1).

[6] R. Sarno and F. Sinaga, "Business Process Anomaly Detection using Ontology-based Process Modeling and Multi-level Class Association Rule Learning," International Conference on Computer, Control, Informatics and its Applications (IC3INA), 2015, pp. 12 – 17.

[7] J. Lathrop and B. Ezell, "A systems approach to risk analysis validation for risk management," Saf Sci, 2017, pp.187-195.

[8] B. Barafort, AL. Mesquida AL, and A. Mas. "Integrating risk management in IT settings from ISO standards and management systems perspectives," Comput Stand Interfaces, 2017, pp. 176-185.

[9] Institute I governance. COBIT 4.1. America: Printed in United States of America; 2007.

[10] A. Joshi, L. Bollen, H. Hassink, S. De Haes, and W. Van Grembergen, "Explaining IT governance disclosure through the constructs of IT governance maturity and IT strategic role," Inf Manag, 2017.

[11] S. Huda, R. Sarno, and T. Ahmad, "Fuzzy MADM Approach for Rating of Process-Based Fraud," J ICT Res Appl, 2015, pp. 111-128.

[12] N. Setiawan and R. Sarno, "Multi-Criteria Decision Making for Selecting Semantic Web Service Considering Variability and Complexity Trade-Off," Journal of Theoretical and Applied Information Technology, vol. 86, no. 2, pp. 316-326, 2016.

[13] BM. Elomda, HA. Hefny, and HA. Hassan, "An extension of fuzzy decisi on maps for multi-criteria decision-making," Egypt Informatics J, 2013, pp. 147-155.

[14] KA. Wiguna, R. Sarno, and NF. Ariyani, "Optimization Solar Farm Site Selection Using Multi Criteria Decision Making Fuzzy Ahp and Promethee: Case Study in Bali," International Conference on Information & Communication Technology and Systems (ICTS), 2016, pp. 237 – 243.

Analysis of Potencies and Investment Opportunities in East Java Based on Geographic Information System

Mas Odi
Information Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
odiek@mhs.mmt.its.ac.id

Riyanarto Sarno, R.V.Hari Ginardi
Departement Of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

***Abstract* - Investment is one of indicators to economic growth and to this effort to ensure sustainable economic growth efforts need to be made investment have been dealt with. At the time there is no a system accurate provides the potential and investment opportunities in a region. So that related department experienced difficulty in attract investor to invest capital and realize investment. The system to be built is Web-based Geographic Information System which provides information on potential and investment opportunities in East Java. The system is combined with Gross Regional Domestic Potential data using Identified Sector Location Location Quotient (LQ) Identification Technique, Growth Ratio Model Analysis, Shift Share Analysis and Overlay Analysis using classifier modeling to know the indication of grouping of sub-regions in an area. In the Kediri district case study the potentials and opportunities of the fishery sector can be seen directly through the geographic information system application.**

Keywords: ***Geographic Information System, Growth Ratio Model Analysis, Location Quotient Analysis, Shift Share Analysis, Overlay Analysis.***

I. INTRODUCTION

Regional potential within the framework of development is defined as the ability, strength and ability in the economic field owned by a region [1]. In the effort of development of a region, the area will utilize potential aspects that can be in the form of natural resource potential, human resource potential, geographical potential and other supporting facilities and infrastructures [2]. Thus, to see the economic potential of a region, it is identified not only natural resources or the ability to procure raw materials and human resources or the ability of the procurement of its workforce, but also aspects affecting other economic productivity, namely geographic potential, social and cultural environment and facilities and infrastructure [3].

Regional potential is an important part of development planning in Indonesia because it can double function in economic growth [4]. First, the economic potential as an input productive activity that encourages economic growth. Secondly, as a medium of increased investment activities of the region which ultimately also affects economic growth. When the potential of the region functions in an increase in capital investment, then it has another multiplier effect such as increased employment, decreased crime rates, poverty reduction, and the like.

Regional potentials as inputs of productive economic activities. In the conceptual economy, natural resources, human resources, technology and other social capital is the input of productive economic activity, namely the activity of producing output both goods and services [5]. BPS classifies productive economic activities into 17 sectors while the Capital Investment Coordinating Board classifies them into 3 main sectors, namely primary, secondary and tertiary sectors. The primary sector group consists of Sectors: 1) Agriculture, and 2) Mining and Quarrying. Secondary sector group consists of Sector: 1) Manufacturing Industry, 2) Electricity & Clean Water, and 3) Building / Construction. The tertiary sector consists of Sectors: 1) Trade, Hotel & Restaurant, 2) Transportation & Communications, 3) Finance, Leasing, & Services Companies, and 4) Services. The twenty-four sectors are engaged in productive economic activities that ultimately produce output in the form of goods and services [6].

II. LITERATURE STUDY

Geographic Information System (GIS) can be defined as an integrated information system for there is a special levy processing the data have the information spatial. In GIS, we can edit the data in the form of maps or specific information, analyze spatial information, maps of the area and present all of the results of operations. In this present era, GIS is a tool that is very useful to analyze the suitability of the land [6]. There are two types of data in a GIS coordinate the data and data raster .Vector data in form points ,a line or polygon .Vector data used to make a map administrative or development plan. Raster data is spatial data that shows the appearance of the earth's surface obtained from remote sensing satellite imagery [7]. SIG complete includes methodology and technology required, the data spatial hardware, software and the organizational structure [8]. GIS is system that could support decision-making spatial and can integrating description location by characteristic of the phenomenon of found in the location [9].

III. METODHOLOGY

In order to promote local investment, the regions need to have information on investment potential and opportunities (investment) from the sector or field of investment business as described previously [10]. This information is in the form of quantitative data of regional potential level through 5 analytical techniques such as 1) LQ analysis; 2) MRP analysis; 3) Overlay Analysis; 4) Shift Share Analysis; 5) Klassen Tipology Analysis. The next five analyzes are described below.

1. Analisa Location Quotient (LQ)

LQ analysis is an analysis used to indicate the size of the role and identify potential economic sector / subsector (base sector), which has comparative advantage in a region. Approach analysis conducted in identifying the leading sector by conducting Location Quotient (LQ) [11]. Location Quotient (LQ) is a technique used to determine the capacity of the regional economy. This technique is used to measure the concentration of an economy in a region by comparing its role in the regional economy with the same economic role in the regional economy, or used to identify the sectors and sub-sectors of activities that have excellent potential [12].

This method is used to identify sectors and sub sectors of economic activity that have potential in Kediri Regency. The formulation of this analytical tool which can be seen in (1):

$$LQ = \frac{S_{ik}/_{PDRB_K}}{S_{iP}/_{PDRB_P}} x\ 100 \qquad (1)$$

Where,

From the results of this LQ calculation can be seen that if:

a. LQ of a sector> 1 is said to be the base sector, since the contribution of sector i in the study area is greater than the contribution of the same sector in the reference area.

b. LQ of a sector <1 is said not the base sector, since the contribution of sector i in the study area is smaller than the contribution of the same sector in the reference area.

c. LQ of a sector = 1 is said to be the provincial level sector level or sector specialization in the study area equal to the same sector in the reference area.

2. Growth Ratio Model Analysis (MRP)

The Growth Ratio Model Analysis (MRP) is an analysis used to identify potential economic sector / subsector based on GDP growth criteria (competitive advantage) [13].

MRP analysis is conducted to mlihat description of economic activities, especially economic structure of Kediri regency that emphasizes the growth criteria both externally (province) and internal (Kediri Regency). The MRP analysis approach can be divided into two, namely: (1) Growth Ratio, and Growth Ratio of Study Area. RPR compares the growth of each activity in the context of East Java province with East Java Province GRDP. Further analysis compares the growth of each activity in the context of the region Kediri Regency with Gross Regional Domestic Growth of Kediri Regency (regional scope). If the RPR value is greater than 1 then RPR is said (+) and if RPR smaller than 1 is said (-). RPR (+)indicates that the growth of a particular activity is at the Provincial or Territory level Kabupaten Kediri higher than growth PDRB Province or PDRB region Kabupaten Kediri. Vice versa if RPR (-). while RPS comparing the growth of activities at the regional level Kabupaten Kediri with the growth of relevant activities at the Provincial or Territory level Kabupaten Kediri. When the growth of an activity at the level of Kediri regency is higher than the growth of the activity at the provincial / district level of Kediri is identified as (+), vice versa if RPS (-). From the Fig. 1 analysis of MRP, will be obtained the real value and nominal value then the combination of the two perbandngan will be obtained description of potential economic activities, namely;

K2 RP_2 (+) RP_3 (-) shows that these activities at the provincial level have prominent growth but at the district level. Kediri not yet prominent	K1 RP_2 (+) RP_3 (+) → dominant growth shows that these activities at the provincial level have a prominent growth and so are at the regional level Kab, Kediri
K4 RPR (-) RPS (-) Show that the activity has low growth at the provincial level and Kabupaten. Kediri	K3 RPR (-) RPS (+) Indicates that the activity is not prominent in the province but at the Kediri Regency level the activity stands out to mean that this activity contributes potentially

Fig. 1. analysis of MRP

3. Overlay Analisys

Overlay analysis is intended to view a description of potential economic activity based on growth criteria and contribution criteria. There are four possibilities in overlay analysis in Fig. 2;

K2 Growth (+) and contribution (-) indicate a sector (subsector) whose growth is dominant but its contribution is small.	K1 Growth (+) and contribution (+) show a sector (subsector) that is very dominant both from growth and from its contribution
K4 Growth (-) and contribution (-) indicate a sector (subsector) that is not potential either growth criteria or its contribution.	K3 Growth (-) and contribution (+) indicate a sector (subsector) whose growth is small but its contribution is large.

Fig. 2. four possibilities in overlay analysis

4. Shift Share Analisys

Shift-Share analysis is an analysis used to identify competitive advantage and know the level of economic specialization in a region.

Shift-Share analysis is a technique that describes the performance (performance) of sectors in a region compared to the performance of the sectors of the national economy. Thus can be found shift (shift) the results of regional economic development, if the region get progress more slowly or faster than the national progress. Previous researchers [14], argued that shift-share analysis is a very useful technique in analyzing

changes in economic structure compared to the national economy. This technique compares the rate of growth of sectors in a region with the rate of growth of the national economy and its sectors, and observes deviations from those comparisons. If the deviation is positive, it is called the competitive advantage of a sector within the region.

There are two equation approaches that can be used to measure the economic growth of a region [14], which can be seen in (2):

$$G = NS + IM + RS \quad \text{or} \quad G = RP + RS \qquad (2)$$

with :

G = Regional Economic Growth = (Eri (t+1) / Eri (t)), to measure the growth of regional gross added value from year to year.

NS = National Share = (En (t+1) / En (t)), to measure national gross value added growth year after year.

IM = Industrial Mix = { (Eni (t+1) / Eni (t)) - (En (t+1) / En (t)) }, to measure the growth of gross national added value of sector i than its total sector.

RS = Regional Shift = { (Eri (t+1) / Eri (t)) - (Eni (t+1) / Eni (t)) }, to measure growth of gross added value of sector i region compared to growth of national gross value added of sector i.

RP = Regional Proportion (RP = NS + IM).

5. Tipology Klassen Analisys

Klassen Tipology is an analysis used to determine the relative potential of sector / sub-sector of Kediri Regency to other districts / cities as East Java Province. In the marticular table of typology klassen, there are two approaches; namely the sectoral approach and the approach of local or regional. Basically Klassen typology method divides the area based on two main indicators, namely regional economic growth and regional per capita income by determining the average economic growth as a vertical axis and average per capita income as a horizontal axis. The observed area can be divided into four classifications in Fig. 3 and Fig. 4:

1) The area is fast forward and fast growing (High growth and high income) is the rate of growth PDRB and per capita income is higher than the average national average percapita growth and income.

2) The area is advanced but depressed. (high income but low growth) ie a relatively advanced region, but in recent years the rate of growth has declined due to the depressed main activities of the region concerned. This area is an area that has been developed but in the future growth will not be so fast although the potential for development is basically very large. This region has higher per capita income but its economic growth rate is lower than the national average.

3) The area is growing fast (high growth but low income) is a fast-growing region with huge potential development potential but not fully processed properly. The level of regional economic growth is very high, but the per capita income level reflecting from the stage of development that has been achieved is actually still relatively low. This area has a high growth rate but percapita income level is lower than the national average.

4) The area is relatively left behind (low growth and low income) is an area that still has a growth rate and percapita income is lower than the national average.

Quadrant I The forward and growing very rapidly $g_i > g_1 \quad gk_i > gk$	Quadrant II Regional forward but depressed $g_i < g_1 \quad gk_i > gk$
Quadrant III he regions that can still be perkembang by leaps and bounds $g_i > g_1 \quad gk_i < gk$	Quadrant IV The relative left behind $g_i < g_1 \quad gk_i < gk$

Fig. 3. Classifications klassen typologies regional approach

As for the sector, picture klassennya typology is as follows:

Quadrant I The forward and growing by leaps and bounds $g_i > g_1 \quad s_i > gk$	Quadrant II Forward but depressed sector $g_i < g_1 \quad s_i > s$
Quadrant III Potential sectors or can still be growing very rapidly $g_i > g_1 \quad s_i < s$	Quadrant IV The relative left behind $g_i < g_1 \quad s_i < s$

Fig. 4. Classifications klassen typologies regional approach for sector

a) advanced and growing sectors (Quadrant I).

This quadrant is a sector quadrant with a growth rate of PDRB (gi) greater than the growth of the reference area or nationally (g) and has a contribution to PDRB (si) greater than the sector's contribution to the PDR regional PFR which is the reference or national (s). This classification is commonly denoted by gi> g and si>s.

b) Sector advanced but depressed (Quadrant II).

The sectors in this quadrant have a lower GDP growth value (GDP) than the GDP growth of the reference region or nationally (g), but contribute to regional GDP (PDRB) greater than the contribution of the sector's value to the regional GDP which is the reference or nationally (s). This classification is usually denoted by gi <g and si>s.

c) Potential sector or still can grow rapidly (Quadrant III).

This quadrant is a quadrant for a sector that has growth value PDRB (gi) which is higher than growth PDRB areas of reference or nationally (g), but the contribution of the sector to PDRB (si) is smaller than the contribution value of the sector to PDRB areas of reference or nationally (s). This classification is commonly denoted by gi>g and si<s.

d) The relatively lagging sectors (Quadrant IV).

This quadrant is occupied by a sector that has a growth value PDRB (gi) which is lower than growth PDRB areas of reference or nationall (g) and at the same time have the contribution to PDRB (si) which is smaller than the value of the sector's contribution to PDRB areas of reference or nationally (s).

IV. EXPERIMENTAL RESULT

1. Level Potency of Investment Regions

1.1 Analisa Location Quotient (LQ)

In the process of economic development planning, local governments prioritize the development and strengthening of sectors in the economic field by developing, improving and utilizing existing resources optimally. Uky [15], stated that within the framework of achieving the objectives of regional economic development, it needs a development policy based on regional peculiarities (endogenous development), using local potential resources, economic development in the region requires special attention and handling.

Each region has different potential with other regions. In the implementation of regional economic development required the existence of appropriate planning and strategy targets, because in every region the situation is not the same. Each region has the characteristics and potential of different areas with each other. The diverse regional potential will shape the structure of the regional economy. Regional economic structure can be seen from the contribution of each sector of the economy to PDRB. By knowing the description of the regional economic structure, the economic development efforts can be more directed in accordance with the potential of the region.

As a first step to identify potential regional potentials, the economic base model can be used. The indicator that an economic sector can be categorized as economic sector is able to become the base sector, that is if it has a Location Quotient (LQ) value greater than one (LQ > 1). Whereas if it has a value less than one (LQ < 1) then belong to the non-base sektoor. Locations Quentient Analysis (LQ) used to know the sectors of the economy in PDRB which can be classified into the bass sector as well as non-base sectors. From calculation based on Location Quotient analysis (LQ) the following results are obtained:

To calculate the value LQ, then the data needed is data PDRB Kabupaten Kediri dan value PDRB East Java Province. Based on the results of the analysis LQ, of the 17 sectors and their sub-sectors based on business fields that play a role in economic activity Kabupaten Kediri, as many as eight sectors that become the leading sectors in the economic movement Kabupaten Kediri. 8 The sectors are 1) Agriculture, Forestry and Fisheries; 2) Large and Retail Trade; Car and Motorcycle Repair; 3) Information and Communication; 4) Real Estate; 5) Mandatory Government, Defense and Social Security Administration; 6) Educational Services; 7) Health Services and Social Activities; 8) Other Services. More can be seen in TABLE I below.

1.2 Results Calculation Analysis LQ

From LQ analysis result from agriculture, forestry and fishery category data from 2011 to 2015 generated average: 1,96 Base, mining category and average quarrying: 0.34 Non Basis, category of processing industry average: 0.66 Non Base, category of electricity and gas procurement average: 0.17 Non Basis, Category of Water Supply, Waste Management, Waste and Recycling averages: 0.57 Non Basis, Average Construction Category: 1.01 Basis, Category of Large and Retail Trade; Average Car and Motorcycle Activity: 1.10 Base, Transportation and Warehouse category on average: 0.56 Non Basis, Category of Accommodation and Drinking Average Drinking: 0.26 Non Base Base, Information and Communication Category on average: 1.05 Base, Average category of Financial Services and Insurance: 0.70 Non Basis, Real Estate category average: 1.22 Base, Category of Service Company average: 0.38 Non Basis, category Administration, Defense and Social Security Compulsory Average: 1.63 Basis, Average educational services category: 1.85 Base, category Health Services and Average Social Activities: 1.26 Basis, Other Service categories on average: 1.26 Basis.

Furthermore, to know the sub-sectors that have competitive advantage then analyzed the Growth Ratio Model (MRP). This data processing uses the average growth PDRB sector in Kediri Regency and East Java Province.

1.3 Analysis results Growth Ratio Model Analysis

From the result of Growth Ratio Model analysis from agriculture, forestry and fishery category data from 2011 until 2015, the growth value 1, the mining category and the 1s growth value growth, the category of growth value 1 industry, the category of electricity procurement and gas the value of growth 1, categories of Water Supply, Waste Management, Waste and Recycling of growth value 1, category of growth value 1, Large and Retail Trade category; Caring and Motorcycles value growth 1, Transportation category and Warehouse growth value 1, category of Accommodation and Feeding Provision Drinking growth value 1, category Information and Communication growth value 1, Financial Services and Insurance growth value 1, Real Estate category growth value 1, category Company Services growth value 1, category Administration Government, Defense and Social Security Mandatory growth value 1, category Education services growth value 1, category Health Services and Social Activities growth value 1, category Other Services value growth 1.

So from several categories of growth Kediri district with the growth of East Java balanced with the decision of one.

1.4 Analysis of Overlay

After doing the analysis LQ and MRP, the analysis is continued by using Overlay analysis which aims to derive a description of potential economic activity in a region based on growth criteria (analysis of the study area or RPs) and contribution criteria (LQ analysis results). According to Kadek [13], there are four possibilities in this analysis: a combination of potential economic sectors or sub-sectors that describe the state of an area as follows:

Overlay analysis results are an analysis used to obtain a description of potential economy based on growth and contribution criteria. Overlay analysis is intended to view a description of potential economic activity based on growth criteria and contribution criteria.

Based on the result of Overlay analysis then obtained, the result in TABLE I below. In TABLE I below shows there are 16 sectors in Kediri Regency indicating that the activities in the

sector is a sector (sub-sector) is very dominant both from growth and from its contribution at the level of Kediri regency. However, there is one sector in Keiri Regency such as electricity and gas procurement sector which is categorized as quadrant 3. This means that the sector of Electricity and Gas Procurement in Kediri Regency is a small economic activity but its contribution is big in Kediri Regency. More can be seen in TABLE I below.

1.5 Result of Overlay Analysis

Result of Overla Analysis from data of agriculture, forestry and fishery category in 2011 until 2015 resulted in growth value 4, mining category and excavation growth value 4, category of processing industry growth value 4, electricity procurement category and gas value growth 3, categories of Water Supply, Waste Management, Waste and Recycling of growth value 3, category of growth value value 3, category of Big and Retail Trade; Car and Motorcycles growth value 3, Transportation category and Warehouse growth value 3, category of Accommodation and Feeding Provision Drinking growth value 3, category Information and Communication growth value 3, Financial Services and Insurance growth value 3, Real Estate category growth value 3, category of Corporate Services growth value 3, category of Government Administration, Defense and Social Security Mandatory growth value 3, category Education Services growth value 3, category Health Services and Social Activities growth value 3, category Other Services growth value 3.

1.6 Shift Share Analysis

Shift-Share Analysis is a technique that describes performance (performance) sectors in a region compared to the performance of the sectors of the national economy. Thus can be found shift the results of regional economic development, if the region get progress more slowly or faster than the national progress. Previous researchers [14], suggests that shift-share analysis is a very useful technique in analyzing changes in regional economic structures compared to the national economy. This technique compares the rate of growth of sectors in a region with the rate of growth of the national economy and its sectors, and observes deviations from those comparisons. If the deviation is positive, it is called the competitive advantage of a sector within the region.

Of these two components if the quantity PS and DS expressed in a plane, with the PS value as the horizontal axis and the DS value as the vertical axis, there will be four categories of relative positions from all regions or sectors of the economy. The four categories are as follows:

• Category I (PS positif and DS positif) is a region / sector with very rapid growth (rapid growth region).

• Category II (PS negatif and DS positif) is the region / sector with the growth rate stunted but growing (depressed region that develops).

• Category III (PS positif and DS negatif) is a region / sector with slowing growth rates but potentially potential (depressed region that potential).

• Category IV (PS negatif and DS negatif) adalah areas or sectors of the depressed region with weak competitiveness and also the role of low-lying areas.

Based on the analysis, from 17 economic driving sector in Kediri Regency, there are 3 sectors included in category 1 of which are 1) Large and Retail Trade sector; car repair and Motorcycles; 2) Transportation and Warehousing sectors; and 3) Information and Communications sector. This means that these three sectors are the sectors with a very rapid growth in Kediri Regency and have great potential for development.

While the sectors included in category 2 of 3 sectors are: 1) Water Supply, Waste Management, Waste and Recycling; 2) Mandatory Government Administration, Defense and Social Security; and 3) Other Services. These three sectors are the sectors with stunted but growing growth rates.

Category 3, there are 8 sectors, including sector 1) Processing Industry; 2) Procurement of Electricity and Gas; 3) Construction; 4) Provision of Accommodation and Drinking; 5) Financial Services and Insurance; 6) Real Estate; 7) Educational Services; 8) Health Services and Social Activities. These eight sectors are sectors with stunted growth speed but potentially potential (depressed regions that potential). Some sectors included in category 4, there are 2 sectors including 1) Agriculture, Forestry and Fisheries; and 2) Mining and Quarrying. Both of these sectors are depressed regions with low competitiveness and also low roles.

1.7 Result Shift Share Analisys

The result of Shift Sharte Analisys analysis from agriculture, forestry and fishery category data from 2011 until 2015 was generated growth value 1, mining category and excavation of growth value 1s, category of industry value growth 1, electricity and gas procurement category 1, categories of Water Supply, Waste Management, Waste and Recycling of growth value 1, category of growth value 1, Large and Retail Trade category; Caring and Motorcycles value growth 1, Transportation category and Warehouse growth value 1, category of Accommodation and Feeding Provision Drinking growth value 1, category Information and Communication growth value 1, Financial Services and Insurance growth value 1, Real Estate category growth value 1, category Company Services growth value 1, category Administration Government, Defense and Social Security Mandatory growth value 1, category Education services growth value 1, category Health Services and Social Activities growth value 1, category Other Services value growth 1. So from several categories of growth Kediri district with the growth of East Java balanced with the decision of one.

Based on some analysis above, among others LQ analysis, MRP analysis, Overlay analysis, Shift Share Analysis and Klassen Tipology Analysis, the Level of Investment Potential in Kediri Regency. Potential Level of Investment Area in Kediri Regency Based on TABLE I above, several sectors / subsectors included in the level of local investment potential are very high among others are sectors or sub-sectors 1)Agriculture, Animal Husbandry, Hunting and Agricultural Services; 2) Construction; 3) Large and Retail Trade; Reaparasi Mobil and Motorcycles; 4)Transportation of land

transportation; 5)Information and Communications; 6)Financial Services; 7)Real Estate; 8)Mandatory Administration of Government, Defense and Social Security; 9) Health Services and Social Activities; 10) Other Services. Then to determine the carrying capacity of investment potential is done qualitative approach which will be discussed further in the next discussion.

2. *Analysis of Offer and Request*

After analyzing the potential of the region, one of the potentials of the region is the sector or sub-sector of Agriculture, Animal Husbandry, Hunting and Agricultural Services. One example to analyze the supply and demand of the case study is the potential of the fishery area is catfish. Following the analysis of supply and demand of catfish are discussed below;

Based on TABLE I total demand catfish production in the year 2016 reached 10,347.00 kg then the value of production is compared with value of market demand of catfish commodity.

TABLE I. Total Demand of Catfish Consumption Kab. Kediri

Average requests per pokdakan	Number of Pokdakan	Estimated demand for catfish for year
201.460 kg	35	7,051,100

Source : Department of Fisheries, Kediri Regency

Based on data of catfish production and demand of catfish, surplus for catfish fish commodity consumption amounted to 3,295,900 kg.Karena commodity value which become area potency have high surplus value then catfish can be offered to investor through some business field which can done from catfish.

Geographic Information System to display potential investment and investment opportunities, Regional Potential Measurement and Measurement of Regional Investment Opportunities in Fig. 5. on the basic data menu explains the Gross Regional Domestic Potential Data (PDRB - ADHB) District and East Java.

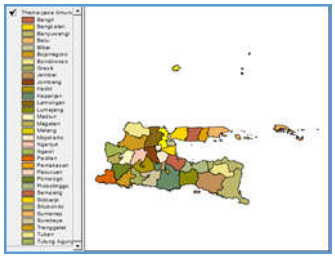


Fig. 5. Screenshot of data of catfish production and demand of catfish

V. CONCLUSION

Web-based Geographic Information System can provide information on potential data and investment opportunities in the Region of East Java in the form of a thematic map of the regions of the sector that is superior and less sectors for investors Domestic and International Investors. From the case study of Kediri regency obtained the potential and the superior opportunity is in fishery sector.

REFERENCES

[1] Y. Charabi, and A. Gastli, "PV site suitability analysis using GIS-based spatial fuzzy multi-criteria evaluation," Renewable Energy,pp. 2554-2561, 2010.

[2] M. Uyan, "GIS-based solar farms site selection using analytic hierarchy process," Renewable and Sustainable Energy Reviews, pp.11-17, 2013.

[3] J. M. Sánchez-Lozano, J. Teruel-Solano, P. L.Soto-Elvira, and M. García-Cascales, "Geographical Information Systems (GIS) and Multi-Criteria Decision Making (MCDM) methods for the evaluationof solar farms locations: Case study in south-eastern Spain,"Renewable and Sustainable Energy Reviews, pp. 544-556, 2013.

[4] J. M. Sánchez-Lozano, C. Henggeler Antunes, M. Socorro García-Cascales, and L. C. Dias, "GIS-based photovoltaic solarfarms site selection using ELECTRE-TRI: Evaluating the case forTorre Pacheco, Murcia, Southeast of Spain," Renewable Energy, pp.478-494, 2013.

[5] K. A. Eldrandaly, and N. Moustafa Abdelaziz, "Enhancing ArcGIS Decision Making Capabilities Using an Intelligent Multicriteria Decision Analysis Toolbox," Journal of EnvironmentalInformatics, pp. 44-57, 2012.

[6] M. Tahri, M. Hakdaoui, and M. Maanan, "The evaluation of solar farm locations applying Geographic Information System and Multi-Criteria Decision-Making methods: Case study in southern Morocco," Renewable and Sustainable Energy Reviews, pp. 1353-1362, 2015.

[7] J. M. Sánchez-Lozano, M.S. García-Cascales, and M.T. Lamata, "GIS based onshore wind farm site selection using Fuzzy Multi-Criteria Decision Making methods. Evaluating the case of Southeastern Spain," Applied Energy, pp. 86-102, 2016.

[8] J. J. W. Watson, and M. D. Hudson, "Regional Scale wind farm and solar farm suitability assessment using GIS-assisted multicriteria evaluation," Landscape and Urban Planning, pp. 20-31, 2015.

[9] H. Selcuk Kilic, S. Zaim, and D. Delen, "Development of a hybrid methodology for ERP system selection: The case of Turkish Airlines," Decision Support Systems, pp. 82-92, 2014.

[10] P. Ghusti, R. Sarno, and R. V. Hari Ginardi. "Substation Placement Optimization Method Using Delaunay Triangulation Algorithm and Voronoi Diagram In East Java Case Study," International Conference on Information & Communication Technology and Systems (ICTS). Pages: 208-213, 2016. DOI: 10.1109/ICTS.2016.7910300.

[11] A. Rikalovic, I. Cosic, and D. Lazarevic, "GIS Based Multi-Criteria Analysis for Industrial Site Selection," in 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013.

[12] U. Yudatama, and R. Sarno, "Evaluation maturity index and risk management for it governance using Fuzzy AHP and Fuzzy TOPSIS (case Study Bank XYZ)," in Intelligent Technology and Its Applications (ISITIA), 2015 International Seminar on, 2015.

[13] K. Aldrin Wiguna, R. Sarno, and N. Fajrin Ariyani. "Optimization Solar Farm Site Selection Using Multi Criteria Decision Making Fuzzy Ahp and Promethee: Case Study in Bali," International Conference on Information & Communication Technology and Systems (ICTS). Pages: 237-243, 2016. DOI: 10.1109/ICTS.2016.7910305.

[14] G. Kabir, and R. Sultana Sumi, "Power substation location selection using fuzzy analytic hierarchy process and PROMETHEE: A case study from Bangladesh," Energy, pp. 717-730, 2014. 242.

[15] U. Yudatama, and R. Sarno. "Priority Determination for Higher Education Strategic Planning Using Balanced Scorecard, FAHP and TOPSIS (Case study: XYZ University)," IOP Conference Series: Materials Science and Engineering (Vol: 105, Issue: 1, Page: 012040). DOI: https://doi.org/10.1088/1757-899x/105/1/012040.

Time and Cost optimization of Business Process Distribution Company using Goal Programming

Gita Intani Budiawati
Department of Management Technology, Faculty of Business and Management Technology
Institut Teknologi Sepuluh November
Surabaya, Indonesia
gitaintani2014@gmail.com

Riyanarto Sarno
Department of informatics, faculty of information and communication of technology
Institut Teknologi Sepuluh November
Surabaya, Indonesia
riyanarto@if.its.ac.id

Abstract— Often when a company is growing and just starting out the management team tend to make all the internal decisions, As the company reaches a certain size Standard operational procedure (SOP) is the important aspect to help the company to evaluate more than what they do now but also help the company to decide how to move forward. Having a standard operating procedure in place streamlines the process and management can accomplish more in optimizing time and cost, on business process of Return Material Authorization (RMA) which may not time and cost effective for the company which was about 2880 minutes with cost IDR. 348,242. In this research, the authors use Goal Programming to optimize time and cost. This approach improves the business process using analysis of multi objective problems. The results show that the time is reduced about 40% and the cost is reduced about 37%, which show improvement in the performance of the business process.

Keywords—optimization; time; cost; standard operational procedure; goal programming;

I. INTRODUCTION

As competition increases and transactions become more complex it is important company's constantly compile data and readjust programs in order to explain and facilitate the work procedure. This will result in increased communication and coordination within all departments of the company. In a company, there must be certain procedures that guide to perform a process [1, 5]. On this paper PT. XYZ is engaged in trading and purchase of wireless devices that great affect the business performance of the company, Therefore it is necessary to analyze the business process to know the existence of non-conformity in the process that is planned with actual performance. While time and cost optimization can be described as a process capable of identifying various activities to obtain minimum duration and minimum cost in a process itself [5, 7, 9].

Establishing SOPs throughout for the company makes it easier to conduct employee performance appraisals. There are some activities in business process of RMA that explore. In business process RMA, the company will replace new goods while it's still under warranty but the company has special rules, if the warranty period passed from less one month, the goods can still be replaced. This special rule likes this often occurs to daily events which have never been described in a previous business process and non-free choice activity drawn to describe this special rule. This paper will contribute to optimizing time and cost for this business process, especially with non-free choice activity.

The model process and simulation will be drawn by Yet Another Workflow Language (YAWL). YAWL is an application used to create and simulate a business process of a company. An event logs of business process simulation is the result of YAWL, which it can use process mining against it [2, 3]. YAWL is used to create a business model process that can be used as a reference or standard in a business process where YAWL can explore the history of every activity that has been done.

Process mining (ProM) helps to analyze business process which include event logs and automatically describes the actual daily process. The goal of process mining analysis is to improve process, not to analyze data from event log [7].

By utilizing the event logs, the business process will be easier to obtain. An event log includes specific records on when and by whom the business activities which support business process [3]. For optimizing cost, the methods used to solve this problem are Goal Programming [9, 8, 13].

II. RESEARCH METHODS

A. Drawing the SOP by YAWL

From the results of interviews and observations, The first step is to describe the SOP in the process business by using Microsoft Visio and drawing it again with YAWL for simulating it, for SOP RMA as for Business process depiction at YAWL and sett time and cost are as follow.

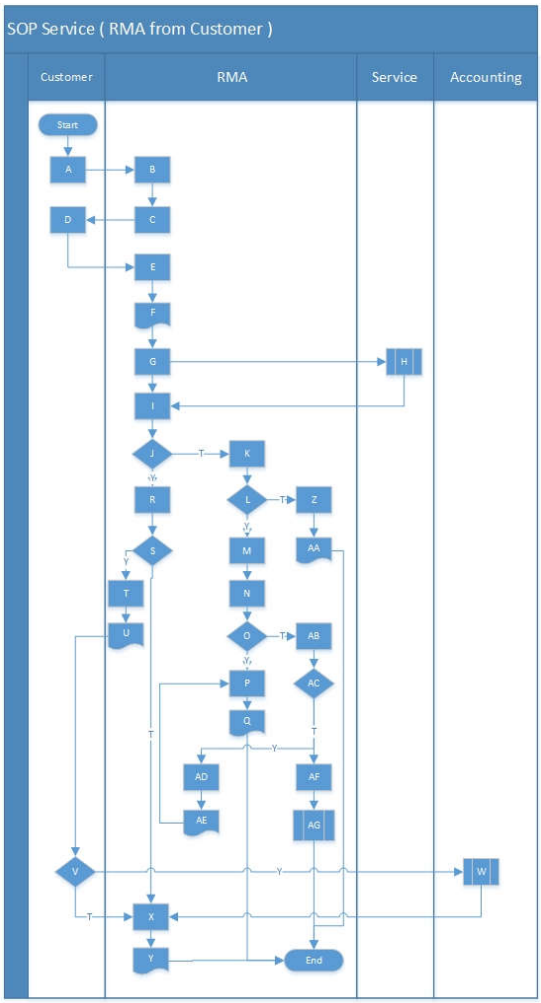


Fig. 1. SOP RMA drawn by Visio

Fig 1 is SOP RMA drawn by Microsoft Visio base of observation and interview. There are 4 departments involved in this procedure. What exactly that they do Table 1 will describe above activities.

TABLE 1. NOTIFICATION OF RMA ACTIVITY

Activities	Note
A	Damaged Goods Information
B	Check Mc.add / Serial Number
C	Info to customer for product warranty
D	Sending damaged goods
E	Create a service receipt
F	RMA Registration Form
G	Submit Goods to Technician
H	Refer To SOP Service
I	Check Service Results from technician via web RMA
J	Results OK?
K	Check Warranty
L	Still guaranteed?
M	Replacement
N	Check RMA Stock
O	RMA Stock Available?
P	Create Job Costing
Q	Job costing form
R	Check if it Is Charge?
S	Is Charge?
T	Create Service Quote
U	Form Repair Service Quote
V	Agree with Quotation?
W	Refer to SOP Billing
X	Create Delivery order RMA (DO)
Y	Form Delivery order RMA (DO)
Z	Create Form Letter notification of damaged goods
AA	Form Letter notification of damaged goods
AB	Check Warehouse Stock
AC	Warehouse Stock Available?
AD	Create form request outgoing
AE	Form Request outgoing
AF	Email Approval purchase
AG	Refer to SOP Purchasing

After drawing using Visio, there are some activities that can be described completely by YAWL. Fig 2 is SOP RMA level 0 where this is the main process of RMA, there are 8 traces and 1 non-free choice activity.

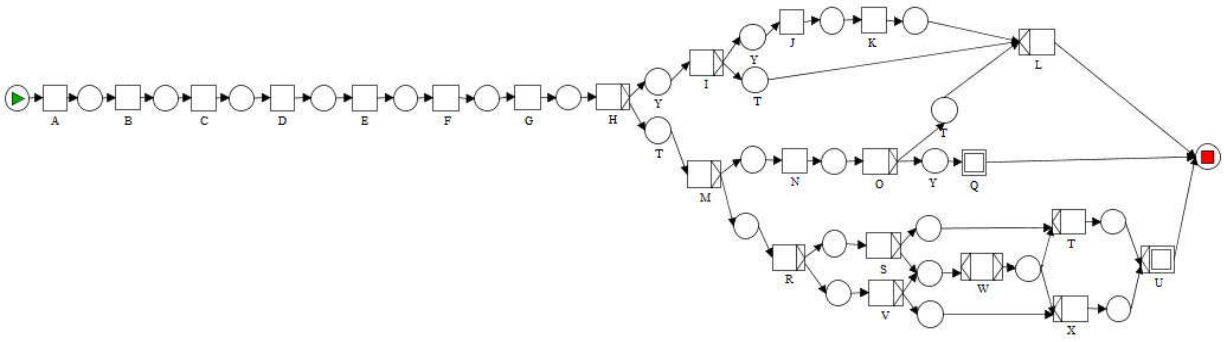


Fig. 2. SOP RMA level 0 drawn by YAWL

As the seen on Fig 2 is SOP RMA level 0 and drawn by YAWL, table 2 will describe the notification of each activity

TABLE 2. NOTIFICATION OF RMA BY YAWL

Activities	Note
A	RMA Request from Customer
B	Check status of goods
C	Info to customer for product warranty
D	Create a service receipt
E	Create RMA Registration Form
F	Submit Goods to Technician
G	Services
H	Check Service Results from technician via web RMA
I	Calculate the charge
J	Create Service Quote
K	Create Billing statement
L	Create Delivery order RMA (DO)
M	Check Warranty
N	Form Letter notification of damaged goods
O	Send Email information to Customer
Q	Send RMA to Vendor
R	Replacement
S	Check RMA Stock
T	Create form request outgoing RMA
U	Create Job Costing
V	Check Warehouse Stock

There are 8 traces and 1 non-free choice activity where on SOP RMA level 0, start from activity *Check RMA stock* then *approval replacement* next activity is non-free choice to *Create form request outgoing RMA* and for activity *Check Warehouse stock* then *approval replacement* next activity is non-free choice to *Create form request outgoing goods*.

Trace 1

RMA Request from Customer- Check status of goods- warranty-reate a service receipt- Create RMA Registration Form- Submit Goods to Technician- Services- Check Service Results from technician via web RMA- Calculate the charge- Create Service Quote- Create Billing statement- Create Delivery order RMA (DO).

Trace 2

RMA Request from Customer- Check status of goods- warranty-reate a service receipt- Create RMA Registration Form- Submit Goods to Technician- Services- Check Service Results from technician via web RMA - Create Delivery order RMA (DO)

Trace 3

RMA Request from Customer- Check status of goods- warranty-reate a service receipt- Create RMA Registration Form- Submit Goods to Technician- Services- Check Service Results from technician via web RMA- Check Warranty- Form Letter notification of damaged goods- Send Email information to Customer- Create Delivery order RMA (DO)

Trace 4

RMA Request from Customer- Check status of goods- warranty-reate a service receipt- Create RMA Registration Form- Submit Goods to Technician- Services- Check Service Results from technician via web RMA- Check Warranty- Form Letter notification of damaged goods- Send RMA to Vendor

Trace 5

RMA Request from Customer- Check status of goods- warranty-reate a service receipt- Create RMA Registration Form- Submit Goods to Technician- Services- Check Service Results from technician via web RMA- Check Warranty- Replacement-Check RMA Stock- Create form request outgoing RMA- Create Job Costing

Trace 6 and non-free choice

RMA Request from Customer- Check status of goods- warranty-reate a service receipt- Create RMA Registration Form- Submit Goods to Technician- Services- Check Service Results from technician via web RMA- Check Warranty- Replacement- Check RMA Stock - Approval Replacement- Create form request outgoing RMA- Create Job Costing

Trace 7

RMA Request from Customer- Check status of goods- warranty-reate a service receipt- Create RMA Registration Form- Submit Goods to Technician- Services- Check Service Results from technician via web RMA- Check Warranty- Replacement- Check Warehouse Stock- Create form request outgoing goods- Create Job Costing

Trace 8 and non-free choice

RMA Request from Customer- Check status of goods- warranty-reate a service receipt- Create RMA Registration Form- Submit Goods to Technician- Services- Check Service Results from technician via web RMA- Check Warranty- Replacement- Check Warehouse Stock - Approval Replacement- Create form request outgoing goods- Create Job Costing

Fig 3 is the level 1 of the create job costing activity, and this process has 1 trace1.

Level 1

Create new form from database – input data customer – choose items – input mc.add – entry data warranty – approve / not approve – print form – request approval – closed job.

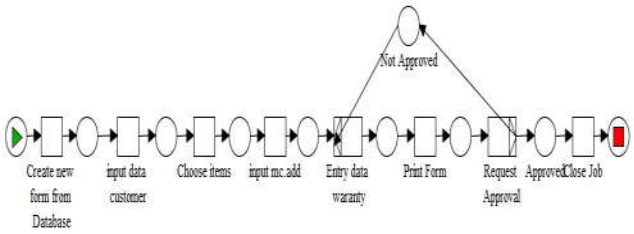


Fig. 3. SOP RMA level 1 Create Job Costing

Fig 4, is the level 1 of Mikrotik Activity, and this process has 2 traces.

Trace 1

Send email to support@mikrotik.com – receive confirmation email – get RMA number – login to mikrotik.com/rma – classification of rejected goods – closing database.

Trace 2

Send email to support@mikrotik.com – receive confirmation email – get RMA number – login to mikrotik.com/rma – Classification of compensation – credit note calculation - closing database

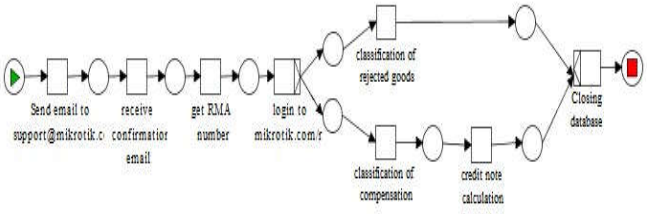


Fig. 4. SOP RMA level 1 of Mikrotik

Fig 5, is the level 1 of Ubiquity, and this process has 1 trace.

Trace 1

Recap items from vendors – login in RMA ubiquity – Entry data – log out.

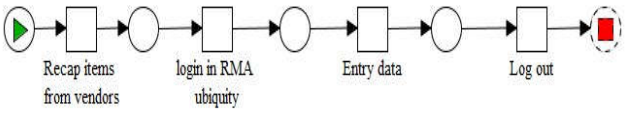


Fig. 5. SOP RMA level 1 of Ubiquity

B. Discovering business process model from even log

The daily routine or daily activities would be written manually based on observation and exporting from the database we can see it on Table 3. There are 4 information Case, Activities, Start Timestamp and resources.

TABLE 3. DATAL OF RMA

Case	Activities	Start Timestamp	Resources
1	RMA Request from Customer	1/31/18 8:00 AM	Requester/Customer
1	Check status of goods	1/31/18 8:04 AM	Admin RMA
1	Info to customer for product warranty	1/31/18 8:06 AM	Admin RMA
1	Create a service receipt	1/31/18 8:10 AM	Admin RMA
1	Create RMA Registration Form	1/31/18 8:18 AM	Admin RMA
1	Submit Goods to Technician	1/31/18 8:23 AM	Admin RMA
1	Services	1/31/18 8:30 AM	Technical
1	Check Service Results from technician via web RMA	1/31/18 10:30 AM	Admin RMA
1	Calculate the charge	1/31/18 10:40 AM	Manager RMA
1	Create Service Quote	1/31/18 11:00 AM	Admin RMA
1	Create Billing statement	1/31/18 11:13 AM	Accounting
1	Create Delivery order RMA (DO)	1/31/18 11:13 AM	Admin RMA
2	RMA Request from Customer	2/1/18 2:00 PM	Requester/Customer
2	Check status of goods	2/1/18 2:04 PM	Admin RMA
2	Info to customer for product warranty	2/1/18 2:10 PM	Admin RMA
2	Create a service receipt	2/1/18 2:13 PM	Admin RMA
2	Create RMA Registration Form	2/1/18 2:17 PM	Admin RMA
2	Submit Goods to Technician	2/1/18 2:23 PM	Admin RMA
2	Services	2/1/18 3:40 PM	Technical
2	Check Service Results from technician via web RMA	2/1/18 4:10 PM	Admin RMA
2	Calculate the charge	2/1/18 4:15 PM	Manager RMA
2	RMA Request from Customer	2/1/18 4:17 PM	Admin RMA
2	Check status of goods	2/1/18 4:30 PM	Admin RMA
2	Create Delivery order RMA (DO)	2/1/18 4:34 PM	Admin RMA

C. Optimizing business process using Goal Programming

Goal Programming is a method used to minimize errors or deviations in doing a goal by considering the priority of the objective [5, 9,10]. Goal programming models are very similar to linear programming models but whereas linear programs have one objective goal programs can have several objectives [8]. A preferred solution is then defined as the one which minimizes the deviations from the set goals. The Goal Programming can be formulated as the following achievement function [13].

$$Min = \sum_{i=0}^{k}(d_i^+ + d_i^-) \qquad (1)$$

$$f(x) + d_i^- - d_i^+ = b_i \quad i = 1, \dots, k, \qquad (2)$$

$$x \subset X, \quad i = 1, \dots, k,$$

$$d_i^- d_i^+ = 0 \quad i = 1, \dots, k,$$

$$d_i^-, d_i^+ \geq 0 \quad i = 1, \dots, k,$$

Where :

x = Number of i in Production
b_i = Level demand for the product i
d_i^- = Deviation of below b_i
d_i^+ = Deviation of above b_i

The purpose of goal programming is to minimize the deviation between the achievement goals (1)(2), $f(x)$ and these acceptable aspiration levels, b_i ($i = 1, \dots, k$,). Also, d_i^-, d_i^+, are respectively, under- and over- achievement of the i goal.

III. RESULT AND ANALYSIS

A. Analyzing the event log with ProM

The result of YAWL simulate are in **.xes** format, the next step is re-extraction into **.mxml** form before those are analyzed by ProM. The processes are analyzed at b on ProM application using alpha++. The information attributes of event log are Case , Activity, Resource and complete timestamp.

TABLE 4. EVENLOG SIMULATE BY YAWL

Case	Activities	Resource	Complete Timestamp
50	RMA Request from Customer	Customer	00:47.1
50	Check status of goods	Nancy	04:13.2
50	Info to customer for product warranty	Nancy	06:19.9
50	Create a service receipt	Nancy	10:41.4
50	Create RMA Registration Form	Nancy	18:09.2
50	Submit Goods to Technician	Nancy	23:14.4
50	Services	Didik	30:14.4
50	Check Service Results from technician via web RMA	Nancy	30:48.8
50	Calculate the charge	Nancy	40:46.6
50	Create Service Quote	Nancy	00:40.3
50	Create Billing statement	Nikita	13:52.5
50	Create Delivery order RMA (DO)	Nancy	13:52.5
51	RMA Request from Customer	Customer	08:47.1
51	Check status of goods	Nancy	11:13.2
51	Info to customer for product warranty	Nancy	16:19.9
51	Create a service receipt	Nancy	20:41.4
51	Create RMA Registration Form	Nancy	28:09.2
51	Submit Goods to Technician	Nancy	33:14.4
51	Services	Didik	40:14.4
51	Check Service Results from technician via web RMA	Nancy	30:48.8
51	Calculate the charge	Nancy	45:46.6
51	RMA Request from Customer	Nancy	05:40.3
51	Check status of goods	Nikita	15:52.5
51	Create Delivery order RMA (DO)	Nancy	17:31.9

B. Analyzing time and cost

The result of optimizing time using ProM analyze by Performance analysis with Petrinet can see in table 5 event log of RMA per log trace

TABLE 5. THROUGHPUT TIME PER LOG TRACE

Log Trace	Throughput time (minutes)	Cost (IDR)
50	193.09	33,660
51	188.75	41,140
52	158.75	37,400
53	158.75	89,760
54	158.75	89,760
55	398.75	119,760
56	151.99	29,920
57	1166.89	41,140
58	1166.89	44,880
59	1165.93	56,100
60	1165.93	41,140
61	1165.93	137,240
62	1159.93	129,760
63	1159.93	129,760
64	1159.93	129,760
65	1159.93	140,980
66	1519.93	129,760

In case 61 until 66 for optimizing time and cost because those case has the longest trace with 20 events than others and also has non-free choice activity.

C. Optimize time and cost using Goal Programming

After analyze optimal of time and choose the traces, the next step is optimizing the cost based on the optimal time. As described above, the optimization of cost will use goal programming method where the longest traces are on case 61 until 66, optimizing cost will also focus on those traces.

Sorce code of Goal programming

```
min = d1plus+d2minus;
137240*C61+129760*C62+129760*C63+129760*C64+140
980*C65+129760*C66+d1plus+d2minus=1000000;
C61>=1;
C62>=1;
C63>=1;
C64>=1;
C65>=1;
C66>=1;
```

Fig. 6. Source code of goal programming

TABLE 6. RESULT OF GOAL PROGRAMMING

Variable	Value	Reduced Cost	Slack or Surplus	Dual Cost
D1Plus	0.0	1.0	0.0	-1.0
D2Minus	0.0	1.0	0.0	0.0
C61	1.0	0.0	0.0	0.0
C62	1.0	0.0	0.0	0.0
C63	1.0	0.0	0.0	0.0
C64	1.0	0.0	0.0	0.0
C65	1.0	0.0	0.0	0.0
C66	2.5	0.0	1.5	0.0

The optimal time and cost on this research are on case 62 (trace 5) with cost IDR129,760 and total time is 1159 minutes also 63 (trace 6) with cost IDR129,760 with total time 1159 minutes. For trace 7, case 65 the optimal time is 1159 minutes with cost IDR140.980 and for trace 8 case 66 the optimal time is 1519 minutes with cost IDR 129,760, as we can see case 66 has value 2.5 with slack or surplus 1.5 it means there are advantages of resource constraints.

IV CONCLUSION

This Paper presented the results of optimizing the business process of a Distribution Company. The optimal traces follow the case number 62 with total time 1159 minute or 19 hours (include office hours) with cost IDR129,760 and case number 63 with total time 1159 minute or 19 hours (include office hours) and cost IDR 129,760. The result shows the improvement of the time and cost before, 40% for time reduced and 37% for cost reduced. It will help companies in improving productivity, as well as cost and time reduction.

REFERENCES

[1] Shabrina Choirunnisa, Riyanarto Sarno, Abd. Charis Fauzan, Optimization of Forecasted Port Container Terminal Performance Using Goal Programming. 2018 International Conference on Information and Communications Technology (ICOIACT),pp. 332-336. 2018.
DOI: 10.1109/ICOIACT.2018.8350719

[2] Lindsay Bradford, Marlon Dums, "Getting started with YAWL,"Handbook, 2007.

[3] Yurika Caesarita, Riyanarto Sarno, Kelly Rossa Sungkono, Identifying bottlenecks and fraud of business process using alpha ++ and heuristic miner algorithms (Case study: CV. Wicaksana Artha). 2017 11th International Conference on Information & Communication Technology and System (ICTS), pp. 143-148. DOI: 10.1109/ICTS.2017.8265660

[4] Yutika Amelia Effendi, Riyanarto Sarno. Non-Linear Optimization of Critical Path Method. 2017 3rd International Conference on Science in Information Technology(ICSITech),pp.90-96.
DOI: 10.1109/ICSITech.2017.8257091.

[5] Shoffi Izza Sabilla, Riyanarto Sarno, Yutika Amelia Effendi. Optimizing Time and Cost using Goal Programming and FMS Scheduling. 2018 International Conference on Information and Communications Technology(ICOIACT),pp.224-249. DOI:10.1109/ICTS.2017.8265660

[6] Riyanarto Sarno, Fitrianing Haryadita, Kartini, Sarwosri, Adhatus Solichah A. Business Process Optimization from Single Timestamp Event Log. 2015 International Conference on Computer, Control, Informatics and its Applications(IC3INA),pp.50-55. DOI:10.1109/IC3INA.2015.7377745

[7] Riyanarto Sarno, Fernandes Sinaga. Business Process Anomaly Detection using Ontology-based Process Modeling and Multi-level Class Association Rule Learning. International Conference on Computer, Control, Informatics and its Applications (IC3INA). Pages: 12 – 17, 2015.
DOI:10.1109/IC3INA.2015.7377738

[8] Nanang Yudi Setiawan, Riyanarto Sarno, Multi-Criteria Decision Making for Selecting Semantic Web Service Considering Variability and Complexity Trade-Off, Journal of Theoretical and Applied Information Technology, vol. 86, no. 2, pp. 316-326, 2016.

[9] Made Agus Putra Subali, Riyanarto Sarno, Yutika Amelia Effendi. Time and Cost Optimization using Fuzzy Goal Programming. 2018 International Conference on Information and Communications Technology(ICOIACT),pp.471-476.
DOI:10.1109/ICOIACT.2018.8350723

[10] Tamiz, M., Jones, D., Romero, C. (1998). Goal programming for decision making: An overview of the current state-of-the-art. European Journal of operational research, 111(3), pp.569-581.

[11] Chowdary, B. V., Slomp, J. (2002). Production planning under dynamic product environment: a multi-objective goal programming approach. University of Groningen

[12] Sharma, S. (2006). Applied Nonlinear Programming. New Delhi : New Age International

[13] IZADIKHAH (2014). Using Goal Programming Method to solve dea problems with value judments in Yugoslav Journal of Operations Research. DOI: 10.2298/YJOR1212210151

Time and Cost Optimization using Goal Programming and Priority Scheduling

Sinarring Azi Laga
Department of Information Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
sinarringazil@gmail.com

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Abstract— The success indicator of an institution and company is when the available resources have been used effectively and efficiently. The limited resources available can cause delays in ongoing business processes. The delay in a business process is always closely related to financing. The slower the business process runs, the more the expenses that must be incurred for labor and material costs. Therefore, it is necessary to optimize the cost and time for an organization. It can be minimized by using goal programming. In goal programming, weights are given to make sense of which activities to be watched. In this research using lingo software to calculate goal programming.

Keywords—Optimization; Schedulling; Goal Programming; Time; Cost

I. INTRODUCTION

In this era of increasingly tight business competition, companies are required to compete in winning market share so that the main objectives of the company can be achieved. One of the company's short-term goals is to make the most profit, while the long-term goal of a company is to maintain survival. The high competition is caused by higher quality and competitive prices of product and substitution products. This condition encourages the company to change the business operation pattern to remain exist. Cost effective in business process is an important issue in optimization problem [1].

This research was conducted at PT.XYZ. PT. XYZ is a company engaged in the field of wireless device sales distribution, where sales activity is the main activity of the company. The accumulation of activity on the sales business process leads to the occurrence of lost time on other business processes. This can impact the profits generated by the company, which can be known from the amount of increasingly bloated expenses incurred for labor and material. The swelling costs are due to improper use of labor. Time– cost enhancement might be characterized as a procedure to distinguish appropriate development exercises to get the base span of process and least extra cost which is required [2]. It can be minimized by using goal programming.

Goal Programming aims to minimize deviations from certain objectives by considering the priority hierarchy [3]. Goal programming can be solved by graphical methods and simplex methods. The simplex method in goal programming is slightly different from the simplex method in linear programming that is the calculation of objective function for each priority. In this research, the optimization of the time performed and the cost incurred in sales activity in PT. XYZ was carried out using goal programming. Goal programming can give an ideal arrangement in cost and time enhancement [4].

II. LITERATURE REVIEW

A. Scheduling

Scheduling is a technique that expects to arrange or apportion leaving assets or machine to play out some assignment inside a specific time span. Scheduling intends to sort out or allot leaving assets or machines to play out some undertaking inside a specific time span. Essentially this is a procedure which decide the action of when it has no begin and stop [5].

B. Priority Scheduling

Priority scheduling algorithm is a scheduling algorithm that prioritizes high priority processes. Each process has its own priorities that have dependencies when executed. In this algorithm, the process that has the highest priority will be run first. Priority scheduling basically has similarities with first come first serve (FCFS) algorithm, but the priority scheduling depends on the highest priority. FCFS is a scheduling algoritm where every ready process will be included in FIFO queue [6]. Here is the formula of the first come first serve algorithm:

$$RT = FT - AT \quad (1)$$

$$TA = \sum RT \quad (2)$$

$$AWT = \frac{TA}{\sum Proses} \quad (3)$$

Where,

RT = Respon Time
FT = Finish Time
AT = Arrival Time
TA = Turn Around
AWT = Average Waiting Time

C. *Goal Programming*

Goal programming is a development of linear programming. Introduced by Charnes and Cooper in the early 1960s [7]. Then, the technique was perfected by Ijiri in the mid 1960s. Thereafter, there was a complete explanation with several development applications by Ignizio and Lee in 1970.

The striking difference between linear programming and goal programming is the structure and the use of the objective function [8]. In linear programming, there is only one goal to be achieved, while there are more than one goal to be achieved in goal programming.

In linear programming, the objective can be either maximization or minimization. While in goal programming, the objective is to minimize deviations from certain goals [9] . This means goal programming is a matter of minimization. There are 3 types of objective function of linear goal programming. The objectives are described below:

1) The objective function in equation 4 is used if the deviation variable in a problem is not distinguished by priority

$$\text{Min } Z = \sum_{i=1}^{m} (d_1^- + d_2^+) \quad (4)$$

2) The objective function in equation 5 is used if the order of objective is required, but the deviation variables within each priority level have the same importance.

$$\text{Min } Z = \sum_{i=1}^{m} P_k(d_1^- + d_2^+);\ k=1,2,..,k \quad (5)$$

3) The objective function in equation 6 is used if the objectives are sorted and the deviation variables at each priority level are differentiated by using different weights.

$$\text{Min } Z = \sum_{i=1}^{m} W_{ki} P_k(d_1^- + d_2^+);\ k=1,2,..,k \quad (6)$$

Where,

m = Number of product or the amount of activity

d_i^- = Deviation variable (presence of destination deviation) below target

d_i^+ = Deviation variable (presence of destination deviation) above target

W_{ki} = The coefficient that represent the contribution to goal k on decision variable i

P_k = Decision variable to k

TABLE I. PROCEDURE FOR ACHIEVING GOAL PROGRAMMING

Minimize	Goal	If Goal is Achieved
d_i^-	Minimize the underachievement	$d_i^- = 0, d_i^+ = 0$
d_i^+	Minimize the overachievement	$d_i^+ = 0, d_i^- = 0$
$d_i^+ + d_i^+$	Minimize the both under-and-over achievement	$d_i^- = 0, d_i^+ = 0$

III. METHODOLOGY

The methodology of this research was arranged systematically as a research framework. Steps of research that will be done, showing in Fig.1.

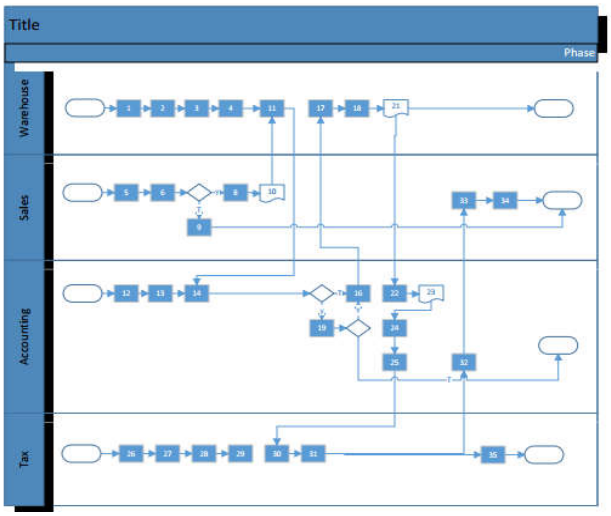


Fig. 1. The proposed method

IV. IMPLEMENTATION

A. *Priority Scheduling*

Based on the priority scheduling algorithm, the first step that must be performed is to determine the priority on the activity [10]. After the activity priority has been determined, then the next step is to calculate the response time, turn around and average waiting time. Here are the priorities on the sales business process:



1. Stuff Come
2. Unloading Stuff
3. Seal of stuff
4. Entry Stuff Database
5. Receive Order
6. Check Availability of Stuff
7. Create Sales Order
8. Customer Info
9. Receive Sales Order
10. Checking Input Balance From B
11. Entry Cash In
12. Check Customer Limit
13. Info Warehouse For Stuff Out
14. Stuff Issued
15. Create Delivery Order
16. Make Billing
17. Create Invoice
18. Entry Database dept
19. Create invoices from tax department
20. Entry Tax Previous Day
21. Make report tax
22. Entry output tax previous day
23. Create an expense tax report
24. Create tax invoice
25. Submit tax invoice to Acc
26. Submit invoices and tax invoices to sales
27. Receive all documents and stuff
28. Handover of documents and stuff to the customer
29. Sending tax invoice online to KPP

Fig. 2. Sales Business Process

Priority scheduling algorithms can be preemptive or non preemptive [11]. If there is a P1 process that arrives when the P0 process is running it will be seen as priority P1, if priority P1> P0 then:

- In Non Preemptive, the algorithm will still complete P0 until its burst time is finished and put P1 in the head queue position.
- In Preemptive, P0 will be stopped first and the process is allocated to P1

Based on the activity in figure 2, it can be known that the arrival time, finish time, response time and average waiting time can be calculated by using priority scheduling algorithm as in table II.

TABLE II. CALCULATION USING PRIORITY SCHEDULLING

No	Activity	Arrival Time	Burst Time	Start Time	Finish Time	Respon Time
1	Stuff Come	0	10	0	10	10
2	Unloading Stuff	1	300	10	310	309
3	Seal of stuff	2	480	310	790	788
4	Entry Stuff Database	3	60	790	850	847
5	Receive Order	4	10	850	860	856
6	Check Availability of Stuff	5	5	860	865	860
7	Create Sales Order	6	15	865	880	874
8	Customer Info	7	5	880	885	878
9	Receive Sales Order	8	5	885	890	882
10	Checking Input Balance From Bank	9	15	890	905	896
11	Entry Cash In	10	25	905	930	920
12	Check Customer Limit	11	6	930	936	925
13	Info Warehouse For Stuff Out	12	2	936	938	926
14	Stuff Issued	13	25	938	963	950
15	Create Delivery Order	14	15	963	978	964
16	Make Billing	15	20	978	998	983
17	Create Invoice	16	15	998	1013	997
18	Entry Database Dept	17	20	1013	1033	1016
19	Create invoices from tax department	18	10	1033	1043	1025
20	Entry Tax Previous Day	19	180	1043	1223	1204
21	Membuat Laporan Pajak Masukan	20	60	1223	1283	1263
22	Entry output tax previous day	21	180	1283	1463	1442
23	Create an expense tax report	22	60	1463	1523	1501
24	Create tax invoice	23	35	1523	1558	1535
25	Submit tax invoice to Acc	24	5	1558	1563	1539
26	Submit invoices and tax invoices to sales	25	5	1563	1568	1543
27	Receive all documents and stuff	26	5	1568	1573	1547
28	Handover of documents and stuff to the customer	27	5	1573	1578	1551
29	Sending tax invoice online to KPP	28	30	1578	1608	1580
					∑Respon Time	**30611**
					Average Waiting Time	**1055.55**

B. *Goal Programming*

The time and cost incurred for the sale activity in this case is taken through the event log, where the unit of time is in minutes and unit of cost is in IDR. The following table is a breakdown of time and expenses incurred on sales activites.

TABLE III. DETAIL OF TIME AND COST USED

	Activity	Time	Cost
1	Stuff Come	10	20960
2	Unloading Stuff	300	462500
3	Seal of stuff	480	495280
4	Entry Database Barang	60	20960
5	Receive Order	10	20960
6	Check Availability of Stuff	5	3460
7	Create Sales Order	15	4960
8	Customer Info	5	1000
9	Receive Sales Order	5	3460
10	Checking Input Balance From Bank	15	21960
11	Entry Cash In	25	3460
12	Check Customer Limit	6	3460
13	Info Warehouse For Stuff Out	2	1000
14	Stuff Issued	25	17500
15	Create Delivery Order	15	4960
16	Make Billing	20	4460
17	Create Invoice	15	22460
18	Entry Database Dept	20	3460
19	Create invoices from tax department	10	4460
20	Entry Tax Previous Day	180	83460
21	Membuat Laporan Pajak Masukan	60	24960
22	Entry output tax previous day	180	83460
23	Create an expense tax report	60	24960
24	Create tax invoice	35	24960
25	Submit tax invoice to Acc	5	5000
26	Submit invoices and tax invoices to sales	5	5000
27	Receive all documents and stuff	5	5000
28	Handover of documents and stuff to the customer	5	17500
29	Sending tax invoice online to KPP	30	23460
	GOAL	**1608**	**1418480**

Here are the variables used in goal programming calculation
X_1 = Stuff Come
X_2 = Unloading Stuff
X_3 = Seal of stuff
X_4 = Entry database barang
X_5= Receive Order
X_6 = Check Availability of Stuff
X_7 = Create Sales Order
X_8 = Customer Info
X_9 = Receive Sales Order
X_{10} = Checking Input Balance From Bank
X_{11} = Entry Cash In
X_{12} = Check Customer Limit
X_{13} =Info Warehouse For Stuff Out
X_{14} = Stuff Issued
X_{15} = Create delivery order
X_{16} = Make Billing
X_{17} = Create Invoice
X_{18} = Entry Database Dept
X_{19} = Create invoices from tax department
X_{20} = Entry Tax Previous Day
X_{21} = Make tax report
X_{22} = Entry output tax previous day
X_{23} = Create an expense tax report
X_{24} = Create tax invoice
X_{25} = Submit tax invoice to Acc
X_{26} = Submit invoices and tax invoices to sales
X_{27} = Receive all documents and stuff
X_{28} = Handover of documents and stuff to the customer
X_{29} = Sending tax invoice online to KPP

The steps for optimization using goal programming are described as follow:

1) Determination of the objective function to be achieved by the company is to minimize the time $Z_{(x)}$. Here is the time minimization equation:
   $\text{Min } Z_{(x)} = 10X_1+300X_2+480X_3+...+30X_{29}$ (7)
2) Determination of the objective function. The objective function to be achieved by the company is to minimize costs. Here is the cost minimization equation:
   $\text{Min } H_{(x)} = 20960X_1+462500X_2+495280X_3+20960X_4\ldots + 23460X_{29}$ (8)
3) Formulation of time limit, Time limits are used according to this following equation:
   $\text{Min } Z_{(x)}=10X_1+300X_2+480X_3+\ldots+30X_{29}\leq 30611$ (9)
4) Formulation of cost limit. Cost limits are used in accordance with this following equation:
   $\text{Min } H_{(x)} = 20960X_1+462500X_2+495280X_3+20960X_4\ldots +23460X_{29} \leq 1608$ (10)

C. *The Result of Goal Programming*

After obtaining goal programming equation based on time and cost, it can be implemented by using lingo software to get optimal time and cost. Here is an input formula on lingo software:

TABLE IV. FORMULA INPUT IN LINGO

Source Code of Goal Programming

```
min Z = D11 + D12 + D21 +D22 ;
20960*x1+462500*x2+495280*x3+...23460*x29-
D11+D12=1418480;
10*x1+300*x2+480*x3+...30*x29-D21+D22=1608;
D11 >=1;
D12 >=1;
D21 >=1;
D22 >=1;
X1 >= 1;
X2 >= 1;
X3 >= 1;
X4 >= 1;
X5 >= 1;
X6 >= 1;
X7 >= 1;
X8 >= 1;
X9 >= 1;
X10 >= 1;
X11 >= 1;
X12 >= 1;
X13 >= 1;
X14 >= 1;
X15 >= 1;
X16 >= 1;
X17 >= 1;
X18 >= 1;
X19 >= 1;
X20 >= 1;
X21 >= 1;
X22 >= 1;
X23 >= 1;
X24 >= 1;
X25 >= 1;
X26 >= 1;
X27 >= 1;
X28 >= 1;
X29 >= 1;
end
```

In goal programming, weights are given to make sense of which activities to be watched. Weight can be feeble 0 or 1 where 0 is as far as possible or still equivalent to mean, while 1 is the farthest point of movement that must exist in the streamlining in light of the fact that the esteem is more noteworthy than anticipated. The outcome parameters of the goal programing are described in table V.

TABLE V. PARAMETERS RESULT OF GOAL PROGRAMMING

Objective Value	Constraint	Variable	d_2^+	d_1^+	d_1^-	d_2^-	Reduced Cost
29007	36	33	29004	0	0	0	29004

Objective value is the most extreme benefit that can be produced. In this case, the object value is 29007. Reduced cost indicates which exercises have a part in producing more benefits. Constraint are restricts on factor esteems. Factors are exercises that keep running on business forms. Here is the result of slack or surplus .

TABLE VI. RESULT OF SLACK OR SURPLUS

Row	Slack Or Surplus	Dual Price
1	29007	-1.000000
2	0.000000	1.000000
3	0.000000	-1.000000
4	0.000000	0.000000
5	0.000000	-2.000000
6	0.000000	-2.000000
7	29003.00	0.000000
8	0.000000	-20950.00
9	0.000000	-462200.0
10	0.000000	-494800.0
11	0.000000	-20900.00
12	0.000000	-20950.00
13	0.000000	-3455.000
14	0.000000	-4945.000
15	0.000000	-995.0000
16	0.000000	-3455.000
17	0.000000	-21945.00
18	0.000000	-3435.000
19	0.000000	-3454.000
20	0.000000	-998.000
21	0.000000	-17475.00
22	0.000000	-4945.000
23	0.000000	-4440.000
24	0.000000	-22445.00
25	0.000000	-3440.000
26	0.000000	-4450.000
27	0.000000	-83280.00
28	0.000000	-24900.00
29	0.000000	-83280.00
30	0.000000	-24900.00
31	0.000000	-24925.00
32	0.000000	-4995.000
33	0.000000	-4995.000
34	0.000000	-4995.000
35	0.000000	-17495.00
36	0.000000	-23430.00

Based on this sales business process, it can be known that the profit generated is 29007 with constraint of X1 to X29 ≥ 1. The Slack or Surplus section in the Solution Report demonstrates how tight the constraint is [12]. On the off chance that a constraint is totally fulfilled as a balance, at that point slack/surplus is zero. In the event that slack/surplus is positive, at that point this tells what number of more units of the variable could be added to the ideal arrangement before the constraint turns into an equity. In the event that slack/surplus is negative, at that point the limitation has been abused. The Dual Price coloumn portrays the sum to which the estimation of the target capacity would enhance if the compelling worth is expanded by one unit.

V. CONCLUSION

The sales case in PT. XYZ has an objective value of 29007 which means it has a very large deviation from the target. It is because the reduced cost for any variable included in the optimal solution value is always zero. The diminished cost for factors barred in the ideal arrangement shows how much the estimation of the target capacity will decline or increment on the off chance that one of the variable units is incorporated into the arrangement. Here is the final result of scheduling result and cost optimization - time that meet the target:

TABLE VII. FINAL RESULT OF SALE ACTIVITY

No	**Activity**	**Constraint Time**	**Burst Time**	**Result**
1	Stuff Come	10	10	Passed
2	Unloading Stuff	300	300	Passed
3	Seal of stuff	480	480	Passed
4	Entry Database Barang	60	60	Passed
5	Receive Order	10	10	Passed
6	Check Availability of Stuff	5	5	Passed
7	Create Sales Order	15	15	Passed
8	Customer Info	5	5	Passed
9	Receive Sales Order	5	5	Passed
10	Checking Input Balance From Bank	15	15	Passed
11	Entry Cash In	25	25	Passed
12	Check Customer Limit	6	6	Passed
13	Info Warehouse For Stuff Out	2	2	Passed
14	Stuff Issued	25	25	Passed
15	Create Delivery Order	15	15	Passed
16	Make Billing	20	20	Passed
17	Create Invoice	15	15	Passed
18	Entry Database Dept	20	20	Passed
19	Create invoices from tax department	10	10	Passed
20	Entry Tax Previous Day	180	180	Passed
21	Membuat Laporan Pajak Masukan	60	60	Passed
22	Entry output tax previous day	180	180	Passed
23	Create an expense tax report	60	60	Passed
24	Create tax invoice	35	35	Passed
25	Submit tax invoice to Acc	5	5	Passed
26	Submit invoices and tax invoices to sales	5	5	Passed
27	Receive all documents and stuff	5	5	Passed
28	Handover of documents and stuff to the customer	5	5	Passed
29	Sending tax invoice online to KPP	30	30	Passed

ACKNOWLEDGMENT

Author might want to express gratitude toward Institut Teknologi Sepuluh Nopember and the Ministry of Research, Technology and Higher Education of Indonesia for supporting the research

REFERENCES

[1] A.T. Rahman, R. Sarno and Y. A. Effendi, "Goal Programming to Optimize Time and Cost for each Activity in Port Container Handling," in International Conference on Information and Communications Technology (ICOIACT), pp. 866-871, 2018. DOI: 10.1109/ICOIACT.2018.8350808

[2] A. Y. Effendi and R. Sarno, "Non-Linear Optimization of Critical Path Method," in *International Conference on Science in Information Technology (ICSITech)*, pp. 90-96, 2017. DOI: 10.1109/ICSITech.2017.8257091

[3] R. Sarno, F. Haryadita, K. S. and S. A. A, "Business Process Optimization from Single Timestamp Event Log," in *2015 International Conference on Computer, Control, Informatics and Its Applications*, pp. 50-55, 2015. DOI : 10.1109/IC3INA.2015.7377745

[4] M. A. P. Subali, R. Sarno and Y. A. Effendi, "Time and Cost Optimization Using Fuzzy Goal Programming," in *International Conference on Information and Communications*, pp. 471-476, 2018. DOI : 10.1109/ICOIACT.2018.8350723

[5] S. I. Sabilla, R. Sarno and A. Y. Effendi, "Optimizing Time and Cost Using Goal Programming and FMS Scheduling," in *International Conference on Information and Communications*, Surabaya, 2018. DOI : 10.1109/ICOIACT.2018.8350727

[6] S. Sharma, Applied Nonlinear Programming, New Delhi: New Age International, 2006.

[7] S. Choirunnisa, R. Sarno and A. C. Fauzah, "Optimization of Forecast Port Terminal Performance Using Goal Programming," in *International Conference on Information and Communications Technology (ICOIACT)*,pp. 332-336, 2018. DOI : 10.1109/ICOIACT.2018.8350719

[8] E. D. Dan and O. D. O, "Goal Programming: - An Application To Budgetary Allocation Of An Institution Of Higher Learning," *Research Journal in Engineering and Applied Sciences ,* vol. 2, pp. 95-105, 2013.

[9] B. Meidyani, R. Sarno and A. L. Nurlaili, "Time and Cost Optimization Using Scheduling Job Shop and Linear Goal Programming Model," in *International Conference on Information and Communications Technology (ICOIACT)*, pp. 555-560, 2018. DOI :10.1109/ICOIACT.2018.8350719

[10] S. Kim, "Efficient exact Boolean schedulability tests for fixed priority preemption threshold scheduling," *The Journal of System and Software,* vol. 134, pp. 153-169, 2017.

[11] R. Begam, Q. Xia and H. Aydin, "Preference-oriented fixed-priority scheduling for periodic real-time task," *Journal of Systems Architecture,* vol. 69, pp. 1-14, 2016.

[12] M. Rizwanullah, K. K and S. K. Verma, "Modeling Of Supply Chain Dynamics: A Lingo Based Three-Tier Distribution Approach," *International Journal of Education & Applied Sciences Research,,* vol. 4, pp. 24-32, 2017.

Vendor Track Record Selection Using Best Worst Method

Ryco Puji Setyono
Department of Information Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ryco.puji@gmail.com

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

***Abstract*— Every company will largely depend on other companies. This will help unite a large business process. Risks that arise from other companies will affect the business performance of a company. Because of this, the right choice for suppliers is crucial. Each vendor has different characteristics. Everything is not always suitable basically the selection process is quite complex and risky. This has led to a new case study which has been studied for years by researchers known as Supplier Selection Problems. Selection of vendors with multi-criteria decision making has been widely studied over years ago. The Best Worst Method is a new science in Multi-Criteria Decision Making (MCDM) determination. In this research, taking case study at XYZ company is in Indonesia which is engaged in mining and industry. The research utilized the transaction data that have been recorded by the XYZ company and analyzed vendor valuation. The weighting of Best Worst Method is calculated based on vendor assessment result. The results show that XYZ company still focuses on Price as its key criteria..**

Keywords- Supplier Selection, MCDM, Best Worst Method, Track Record.

I. INTRODUCTION

Now information is rapidly spreading. Choosing an information is not arbitrary. It must be in accordance with needs. This is the basis of a company in choosing vendors and suppliers as partners. With a large enough information flow, companies can get information about vendors. This information can come from sources outside the company or from internal companies. Application of multiple criteria in assessing vendor performance in various aspects is a necessity [1].

This assessment is needed because there are many existing vendors with various characteristics. Each vendor has different characteristics, performance and capabilities. All the risks posed by the vendor become a company risk as well when become a partner. This becomes an external risk beyond the control of the company. [2]

The research was conducted at the XYZ company engaged in mining and mining industries. Where the business scope is wide enough that requires different item specifications. It becomes quite interesting where every scope of the business process requires different items and goods. Multiple criteria become mandatory to assess this to suit the needs of the company. The complexity of the assessment criteria will be complex. Not only things that are tangible such as price and delivery time, but also things that are intangible such as attitude, vendor commitment and relationships between companies [3].

Vendor selection is an important activity in the procurement department. This can be related to the quality of production to the continuity of business processes. Risk mitigation is important in the procurement process. The right strategy in selecting vendors is the emphasis in this matter. There is no best strategy for vendor selection, but many approaches can be used. [4]. Most of the business processes in industry are raw material production, so vendor selection is vital. The costs incurred for the selection of raw materials can reach up to 70% of production costs. With this situation the division of employment is a key role in reducing costs and mitigating risks. This makes supplier selection one of the most important functions of purchasing management [5].

The key criteria on latest research on selection of vendors and suppliers that most picked are product quality, delivery, history and warranty policy [4]. This research was conducted at XYZ company located in Indonesia. It will concluded the key criteria of valuation from XYZ company. Criteria are the main focus in the vendors and suppliers selection.

The XYZ company has recorded the vendors that work together. The data is stored in the database as an archive used for financial reporting purposes. This data is an internal source of company information. This information is in the form of all transactions that have been carried out with vendors and how the vendors treated them during the same work as information on delivery dates, quality of goods and the prices. There is also data from the external source of the company. The data in the form of information can be obtained from the vendor's reputation and information from the company's partners. This information is processed internally to be used as a collection of information that is useful in selecting vendors.

II. LITERATURE REVIEW

A. Past Research

During the last few decades of research on supplier selection problem, the researcher produce considerable

literacy with various methods used. So that the process of selecting suppliers becomes a quite crucial factor. Especially in the era of globalization, where companies are required to continue to improve competence [6].

In the selection of vendors, the procurement division is taking an election strategy. Where this election is tailored to the needs and capabilities of the company. The selection of vendors in the procurement division will take on the role of Decision Maker. Selected vendors will influence the profitability of the company and customer satisfaction obtained [7].

For a long time, many studies have investigated techniques, models and methods for decision making for vendor selection [8]. There are no best models and techniques because each company has different treatment. Only suitable ones can be applied in the company. Several studies have introduced several methods. A slight change in the vendor selection process can have a significant effect on the supply chain [9]. Some research conduct to prove Multi Criteria Decision Making can use many method that fit with the context.

Research on supplier selection with lexicographic goal programming method and Analytic Hierarchy Process in vendor selection has been done by Cebi and Bayraktar [10]. Kwang et al chose to use scoring method and fuzzy logic in conducting vendor selection research [11]. Research conducted by Degraeve et al by analyzing the total cost of ownership information to evaluate strategy in decision making at European Metal Company [12]. Research about MCDM by developing web applications using Analytic Hierarchy Process method was done by Akarte et al [13].

The Ghodsipour and O'Brien studies used the Analytic Hierarchy Process method and Linear Programming methods to take into account tangible and intangible factors when selecting suppliers [14]. The use of Fuzzy Analytic Hierarchy Process in determining preferred vendors has also been done by Cengiz et al [15]. The incorporation of Taguchi's loss function method and Analytic Hierarchy Process in determining supplier selection has been introduced by Pi and Low [16]. Rao designed a method for vendor selection by combining Genetic Algorithm with Analytic Hierarchy Process [17]. Yadav and Sharma introduced three alternative selection of vendors based on Analytic Hierarchy Process [18]. Amid et al conducted the study by suggesting using a linear multi objective model [19].

B. Criteria of Selection

Supplier selection is quite complex where many criteria need to be considered. Not only tangible criteria but also intangible criteria. Weber et al stated that many studies with various approaches to vendor selection have also been conducted since the 1960's. There are 74 articles covering vendor selection from 1960 to 1991 [20].

The study by Dickson et al in 1966 has stated 23 attributes in the supplier selection criteria. The research conducted by Dickson by reviewing 273 questionnaires distributed to companies in the United States and Canada [21]. The results of the study can be seen in table 1. The criteria offered by Dickson include the criteria for tangible and intangible. There are many factors influence the Supply Chain Environment. Weber et al examined the results of Dickson et al. of and from the 23 criteria examined, at least 1 key criterion from Dickson was mentioned from 74 articles interviewed by Weber et al.

TABLE I. DICKSON'S CRITERIA

No	Criteria	Mean
1	Quality	3,508
2	Delivery	3,417
3	Performance History	3,998
4	Warranties & Claims Policies	2,849
5	Production Facilities and Capacity	2,775
6	Price	2,758
7	Technical Capability	2,545
8	Financial Position	2,514
9	Procedural Compliance	2,426
10	Communication Systems	2,426
11	Reputation and Position in Industry	2,412
12	Desire for Business	2,256
13	Management and Organization	2,216
14	Operating Controls	2,211
15	Repair Service	2,187
16	Attitude	2,120
17	Impression	2,054
18	Packaging Ability	2,009
19	Labor Relations Record	2,003
20	Geographical Location	1,872
21	Amount of Past Business	1,597
22	Training Aids	1,537
23	Reciprocal Arrangement	0,610

As time goes by, the criteria proposed by Dickson began to experience a shift. Not all criteria proposed by Dickson are in line with current Supply Chain conditions. In Table II, Cheraghi, Dadashzadeh et al has conducted a study of Dickson's criteria correlation with research conducted from 1966 to 2001. Some of the criteria that Dickson proposed experienced a shift in values.

In the range of research conducted by Cheraghi et al [22] there were many significant changes. Packaging Ability criteria which after 1990 were no longer as criterion in the study. Because the rapid growth of geographic location transportation and communication technology now is less important. What remains consistent is the criteria are technical capability. Because the need for a company's ability to handle orders is important enough to reduce the risks that can be caused.

With the rapid development of information, customers increasingly understand the products provided by suppliers or vendors. The customer knows how the product specifications are provided, the terms and conditions of a product. After-sales service is one of the company's wishes in selecting suppliers or vendors. The trend of assessment criteria began to focus on customer satisfaction. By gaining greater emphasis, the criteria for sales repair service experienced a rapid increase [22].

C. Modified Criteria for Internal Business Company XYZ

Procurement Division faces a problem where the assessment is tangible and intangible. Criteria adopted for the study derive from the modification of research owned by Dickson et al [21] with consideration of stakeholders and shareholders of XYZ companies. Criteria taken based on

consideration of the company's business process for the procurement division. Business strategy XYZ company based on preference of shareholders will be criteria as well. From these criteria are divided into criteria and sub criteria that can be seen in table III.

Most of these assessments at XYZ company has been enclosed in internal applications. The data and information on vendor valuations are used directly extracted from there.

Reputation	History Transaction
	Current Position in Market
	Partner
Quality Systems	Management
	Assurance
	Waranty
	Production Quality
Price	Discount
	Delivery Cost
	Term of Payments

TABLE II. SUPPLIER CRITERIA BETWEEN THE PERIODS OF 1996 – 1990 AND 1990 – 2001 BY CGERAGHI, DADASHZADEH ET ALL

Criteria	1966 - 1990		1990 - 2001		Overall	
	Papers	%	Papers	%	Papers	%
Quality	40	54	31	79	71	63
Delivery	45	61	30	77	75	66
Performance History	7	9	4	10	11	10
Warranties & Claims Policies	1	1	0	0	1	1
Production Facilities and Capacity	25	34	10	26	35	31
Price	55	74	26	67	81	72
Technical Capability	19	26	11	28	30	27
Financial Position	8	11	7	18	15	13
Procedural Conpliance	2	3	2	5	4	4
Communication Systems	3	4	4	10	7	6
Reputation and Position in Industry	9	12	1	3	10	9
Desire for Bussines	2	3	0	0	2	2
Managemen and Organization	10	14	7	18	17	15
Operating Controls	5	7	0	0	5	4
Repair Service	7	9	11	28	18	16
Attitude	9	12	5	13	14	12
Impression	4	5	2	5	6	5
Packaging Ability	5	7	0	0	5	4
Labor Relations Record	3	4	1	3	4	4
Geograpichal Location	15	20	2	5	17	15
Amount of Past Business	1	1	0	0	1	1
Training Aids	3	4	0	0	3	3
Reciprocal Arragements	3	4	2	5	5	4

Furthermore, for the assessment of criteria, a joint discussion with shareholders and stakeholders will be carried out. The results of the discussion will be included in the Best Worst Method model to get weighting. From there, you can get the results of weighting for XYZ companies in the selection of vendors, as well as the key criteria.

TABLE III. MODIFIED CRITERIA FOR XYZ VENDOR SELECTION

Criteria	Sub Criteria
Performance	Shipment
	Delivery
	Service and Communication
Technical Capability	Technical Cooperation
	Employee Profile
	Equipment
	Manufacturing
	Organization Culture

III. METHODOLOGY

There have been many studies using Multi-criteria concept, in some previous studies using AHP and SAW methods to calculate Risk Analysis [23]. Research Risk Analysis, assessing each risk in one particular criterion. In the Multi Criteria study for semantic web service selection using many criteria in assessing websites, where these criteria are connected and have their own weight [24]. The selection of singers using the AHP and SAW approaches also uses criteria with different weights [25]. Many studies with multi criteria that have been described previously, so the selection of vendors using that concept is very possible.

From many methods used in various studies using MCDM [26], this research uses the Best Worst Method (BWM) developed by Rezaei [27]. This method compares the criterion

criteria that become references. The Best Worst Method uses a bit of data in its assessment, produces consistent data and does not require a full pairwise comparison matrix. This method has been used in water scarcity management research [28], research into measuring efficiency of unicersity-industry projects [29], research in evaluation scientific output [30], research in RnD department evaluations [31] and various other studies.

Here is the steps of BWM.

The first step is to determine the criteria that will be weighed and compared.

The next step, these criteria are selected which are more desirable or we call Best, and which are less desirable, we call Worst. Then each criterion used is determined by using a scale of 9 (numbers between 1 and 9; 1: B is equally important to j; 9: B is extremely more important than j). The result is a best-to-others (BO) vector as follows.

$A_B = (a_{B1}\ a_{B2}\ a_{B3}\ ....,\ a_{Bn})$, with a_{BJ} represent the preference of B over j and $a_{bb} = 1$.

After that, we also define the opposite of the worst criterion (W), using a 9-point scale (numbers between 1 and 9; 1: j is equally important to W; 9: j is extremely more important than W), which results in the others-to-worst (OW) vector as follows.

$A_w = (a_{1w}, a_{2w}, a_{3w}, \ldots, a_{nw})^T$, with a_{JW} the preference of j over W and $a_{WW} = 1$.

This step needs to be input from an expert to determine the value of each criterion. During this research, the expert was a stakeholder and shareholder of the XYZ company. Determination of the Best and Worst criterion, the researchers conducted interviews and meetings. We did not explicitly mention it when in collecting data.

After getting the assessment, the next step is finding optimal weights ($w_1^*, w_2^*, \ldots w_n^*$). To get optimal weigth need maximum absolute differences $\{|w_B - a_{Bj}w_j|, |w_j - a_{jw}w_w|\}$ for all j minimized.

$$\sum_j w_j = 1$$

$w_j \geq 0$, for all j. …(1)

Problem (2) is equal to the following linear problem: min ξ^L.

$$|w_B - a_{bj}w_j| \leq \xi^L, \text{ for all j}$$
$$|w_j - a_{jw}w_w| \leq \xi^L, \text{ for all j} \quad \ldots(2)$$

$$\sum_j w_j = 1$$

$w_j \geq 0$, for all j. …(3)

Solving problem (3), we will get optimal weights (w1*, w2*, … wn*) and optimal value ξL, which is the consistency index. When the value is close to zero, it has a high level of consistency of the pairwise comparisons provided by the decision-maker (s) / expert (s). For problems with more than one level, each level is calculated accordingly. After each level has its own weight, then to get global weight, multiplication is calculated from each level.

IV. Analysis and Discusion

We hold meetings with stakeholders and shareholders XYZ company for five times without discussing directly about vendor valuation. The criteria are adjusted to the needs of the company. Results of the next weighting will be processed to produce data as in table IV. In Figure I, Price is the top priority compared to the main criteria. Undeniably, the supply chain climate conditions in Indonesia still prioritize prices.

The next sequence is Quality Systems, which are criteria related to the quality of an item. The lowest is the vendor's reputation criteria according to the management in the procurement division. Reputation is important, but there are other important things such as vendor service performance and technical capability of the vendor. Even though reputation is important, if the service is bad the risk will be large. The historical data obtained in the XYZ company becomes a reference if it is going to make a transaction.

TABLE IV. Weighted Criteria for XYZ Vendor Selection

Criteria	Weight Criteria	Sub Criteria	Sub Weight Criteria	Global Weight Criteria
Performance	0.1898	Shipment	0.4579	0.0869
		Delivery	0.3850	0.0731
		Service and Communication	0.1576	0.0299
Technical Capability	0.1498	Technical Cooperation	0.2783	0.0417
		Employee Profile	0.1649	0.0247
		Equipment	0.1695	0.0254
		Manufacturing	0.1147	0.0172
		Organization Culture	0.2723	0.0408
Reputation	0.1049	History Transaction	0.3167	0.0332
		Current Position in Market	0.4874	0.0511
		Partner	0.1961	0.0206
Quality Systems	0.2553	Management	0.1362	0.0348
		Assurance	0.2314	0.0591
		Warranty	0.2996	0.0765
		Production Quality	0.3334	0.0851
Price	0.3002	Discount	0.4689	0.1407
		Delivery Cost	0.1743	0.0523
		Term of Payments	0.3568	0.1071



Fig. 1. Weight Criteria on company XYZ.

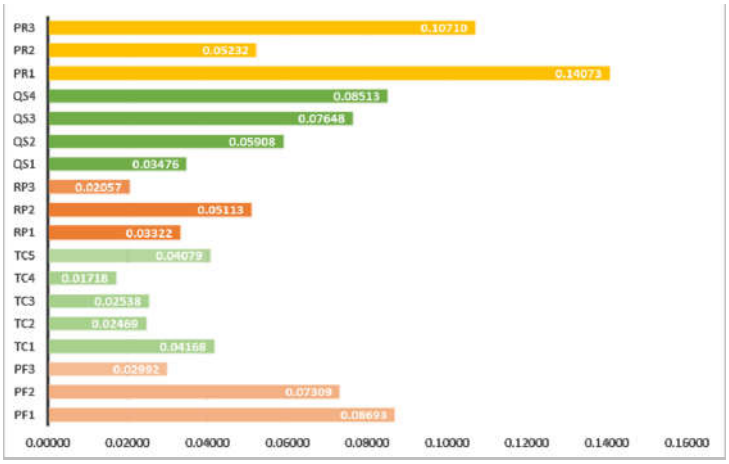


Figure 2. Weighted Sub Criteria from global weight.
- ***Price perspective: pr1:*** Discount, ***pr2*** : Term of Payment, ***pr3*** : Delivery Cost.
- ***Quality Systems : qs1*** : Management, ***qs2*** : Assurance, ***qs3 :*** Warranty, ***qs4:*** Production Quality.
- ***Reputation : rp1*** : partner, ***rp2***: history transaction, ***rp3***: current position in market.
- ***Technical Capability: tc1*** : Organization culture, ***tc2***: manufacturing, ***tc3***: equipment, ***tc4***: Employee profile, ***tc5:*** Technical Cooperation.
- ***Performance : pf1*** : Shipment, ***pf2*** : Delivery, ***pf3*** : Service and Communication.

Vendor data used is the result of transactions with vendors stored in the procurement application. To complete vendor data, general data collection is carried out. Researchers and procurement divisions collect for criteria that are not covered by the procurement application. Such data, such as vendor positions in the market, employee profile sub-criteria and others criteria can be obtained from the website of each vendor. If there is none can be found the procurement division provides an assessment. The assessment is directly ask a sample of employees from vendors who have interacted. The Organization culture is assessed from the data that was obtained when registering a vendor. The vendor registration business process must include data on the vision and mission of the vendor. For data management, it can be taken from the SOP pattern of the vendor, whether it is complicated or not. When making direct transactions with the vendor it can be assess by procurement team as data management. Shipment and delivery data can be seen from historical data when making transactions. Service and communication sub criteria can be obtained from the experience of the procurement division in conducting transactions and communicating with vendors.

This is efficient because the procurement division is quite proficient in assessing vendors. In addition, the procurement division understands the field they are facing, understands vendor behavior that is problematic, and how to react. Assisted with the data that has been stored in the procurement application, vendor assessment can be fairly easy to do. The result can be view at table V.

Sub criteria can be seen in figure 2 where all sub criteria have become global weight. It can be seen that two sub-criteria of Price dominate, namely the discount sub-criteria and the Terms of Payment sub-criteria. Both of these criteria are still a weight for XYZ companies, which can help in increasing the price difference that helps in suppressing production costs and adding higher margins.

The vendor score data with the results of weighting can be seen in table V. There is no vendor that dominates. There must be something less than the perspective of the company. This can also be used in determining vendors in the future. This valuation is completely unknown to the vendor. This data is a confidential company. Except procurement team know about it.

From these results, the vendors with the lowest prices have poor quality. But there are vendors with the best quality and the prices are quite significant with qualified services. From the vendor's point of view, to increase the data on the quality of services and goods requires more costs at the expense of several aspects of the product.

In some vendors there are those whose value is quite low compared to other vendors. There is still a possibility that the vendor has several aspects that still need to be improved and are in an effort to develop it. Also the possibility that the vendor is still improving itself in providing its products and services. Does not close the possibility that the vendor is not focused on market segmentation for the mining industry. It can be the vendor is chosen as a supporting or complementary in the XYZ company business process.

The vendor rating for each criterion can be seen in Table V. Of the 374 vendors listed on the XYZ company system, only 70 samples are taken. The samples are drawn at random. Furthermore, the result of the valuation, multiplied by the weighting that has been obtained from table IV. The multiplication result by weighting the criteria of bias is seen in table VI, there also provided the rating.

TABLE V. VENDOR SCORING FROM COMPANY XYZ DATABASE

Firm No	Performance			Technical Capability					Reputation			Quality Systems				Price		
	pf1	pf2	pf3	tc1	tc2	tc3	tc4	tc5	rp1	rp2	rp3	qs1	qs2	qs3	qs4	pr1	pr2	pr3
1	8	6	8	7	7	6	5	9	9	7	8	6	8	7	8	5	6	6
2	7	7	8	8	6	5	4	8	9	4	8	9	6	7	7	8	7	8
3	6	5	7	6	8	7	9	4	7	7	9	6	5	6	8	7	6	7
4	7	7	6	8	4	7	7	7	7	5	5	6	8	4	7	8	6	7
5	6	5	7	7	6	5	6	8	9	8	7	8	7	7	7	8	7	8
6	6	6	6	5	7	4	7	6	6	7	7	6	5	7	5	1	6	3
7	7	7	7	7	7	8	7	9	7	9	7	7	4	7	6	7	5	6
8	8	6	6	8	6	6	6	6	6	8	7	5	4	7	8	7	6	8
9	8	6	7	7	6	9	6	6	6	8	8	6	7	5	8	5	5	6
10	9	7	7	4	7	7	8	7	5	6	4	7	8	7	9	7	6	8
11	7	7	6	9	5	5	7	8	7	7	4	9	8	7	8	5	8	7
12	7	4	6	7	6	8	9	5	8	9	6	5	7	5	8	6	7	8
13	7	8	6	8	5	6	7	7	7	7	7	7	6	6	6	8	5	8
14	6	6	6	8	6	6	6	6	6	6	7	9	6	8	7	9	6	7
15	5	7	7	7	6	7	7	8	7	8	8	7	9	7	6	5	7	6
16	8	6	8	8	9	7	6	6	8	8	8	7	4	6	8	8	7	5
17	5	7	6	8	8	6	7	6	6	9	8	6	7	6	6	6	7	8
18	4	9	6	5	6	7	8	6	5	5	7	6	7	7	7	4	6	7
19	6	7	4	8	5	8	6	7	8	9	6	7	6	7	6	6	8	6
20	8	9	6	9	6	6	6	8	7	6	8	7	6	8	7	8	6	7
21	4	3	5	7	5	4	7	6	4	7	5	5	6	6	8	6	8	7
22	5	5	6	6	5	4	5	4	6	5	6	4	3	7	3	8	6	9
23	6	5	4	5	6	5	6	6	6	8	5	6	7	9	5	4	6	2
24	4	5	7	6	6	5	4	5	6	4	6	5	6	4	6	6	4	7
25	4	2	5	5	5	5	2	4	7	8	5	6	5	5	6	5	4	9
26	8	8	6	5	3	4	5	2	6	6	4	5	9	6	7	4	8	5
27	8	8	8	9	8	7	6	8	7	8	6	7	8	8	7	7	9	9
28	7	9	7	8	7	8	9	7	7	8	6	7	8	8	7	9	8	6
29	7	8	8	8	5	8	9	8	8	9	9	7	8	7	8	6	7	8
30	8	6	7	9	7	8	8	7	7	7	8	8	6	8	8	5	8	9

TABLE VI. VENDOR SCORING WITH WEIGHTED CRITERIA

Firm No	Performance		Technical Capability		Reputation		Quality Systems		Price	
	Total	Rank	Total	Rank	Total	Rank	Total	Rank	Total	Rank
1	1.3733	17	1.0699	19	0.8214	13	1.8975	17	1.6602	60
2	1.3595	20	1.0036	29	0.6680	38	1.7985	23	2.3489	8
3	1.0964	46	0.9431	34	0.7755	19	1.6438	37	2.0488	33

4	1.2996	26	1.0157	26	0.5910	51	1.5830	42	2.1895	21
5	1.0964	47	0.9962	31	0.8519	6	1.8229	19	2.3489	6
6	1.1396	40	0.8478	39	0.7011	34	1.4649	50	0.7760	70
7	1.3295	24	1.1550	11	0.8366	11	1.5257	46	1.8893	45
8	1.3135	25	0.9817	32	0.7523	25	1.6265	38	2.1559	23
9	1.3434	22	1.0162	25	0.7728	20	1.6855	32	1.6079	63
10	1.5034	7	0.9402	35	0.5551	60	2.0174	4	2.1559	22
11	1.2996	27	1.0721	18	0.6727	36	2.0018	6	1.8719	48
12	1.0804	49	1.0016	30	0.8493	8	1.6508	34	2.0675	31
13	1.3727	19	1.0150	27	0.7344	30	1.5674	44	2.2443	17
14	1.1396	41	0.9817	33	0.6500	39	1.8750	18	2.3302	10
15	1.1557	37	1.0642	21	0.8060	14	1.8211	20	1.7125	57
16	1.3733	18	1.0812	17	0.8393	10	1.6195	39	2.0276	36
17	1.1258	42	1.0483	23	0.8240	12	1.5917	41	2.0675	32
18	1.1850	36	0.9164	36	0.5657	59	1.7533	27	1.6266	62
19	1.1529	39	1.0486	22	0.8493	7	1.6439	35	1.9056	44
20	1.5327	5	1.1050	16	0.7038	32	1.8055	21	2.1895	20
21	0.7166	69	0.8818	38	0.5936	49	1.6682	33	2.0127	38
22	0.9796	54	0.7241	55	0.5783	57	1.1070	68	2.4037	4
23	1.0067	52	0.8313	41	0.7111	31	1.7360	29	1.0911	69
24	0.9226	59	0.7978	44	0.5272	63	1.3449	59	1.8034	50
25	0.6435	70	0.6563	59	0.7443	27	1.3971	53	1.8769	47
26	1.4596	11	0.5515	69	0.5883	54	1.7603	26	1.5170	65
27	1.5195	6	1.1798	7	0.7649	22	1.9237	13	2.4199	3
28	1.4757	10	1.1495	12	0.7649	23	1.9237	14	2.3278	11
29	1.4326	14	1.1409	13	0.9110	1	1.9323	10	2.0675	30
30	1.3434	23	1.1740	8	0.7549	24	1.9254	12	2.0862	28

V. CONCLUSION

Study to measure weighting with multiple criteria has been done a few decades ago. The method of retrieving data from transaction history, survey data from various sources and process it with BWM is very handy in measuring vendor evaluation values. This value is a representative of the XYZ company with the assessment is entirely based on the angle of the company. And assessment held without notification to others party. Assessing this main point of view is enough to be a reference in weighting and evaluating vendors.

As the result from BWM, the criteria that are top priority for XYZ company are Price and discount as sub-criteria. Prices are a priority for XYZ companies and this can be a representative for others Indonesia company. For quality criteria in second priority after the price. The service quality criteria or products occupy the third position on the global weight criteria. Followed by shipping criteria. This is quite reasonable considering the geographical conditions in Indonesia which are quite challenging. Sometimes there are vendors with less shipping capabilities due to geographical conditions. This will hamper business processes from XYZ companies if they experience delays in delivery. With assessments assisted by BWM it is very helpful in evaluating vendors with various criteria. The results were also received by stakeholders and shareholders when disclosing the results.

REFERENCES

[1] Liao, Z. and J. Rittscher. "A Multi-objective supplier selection model under stochastic deman conditions." *International Jurnal of Production Economics,* vol 105, (2007) page: 150-159

[2] Christopher M. "Creating Resilent Supply Chains." *Logistics Europe.* (2004) page: 14-21.

[3] Chai, J. J. N. K Liu and E. W. T. Ngai. "Application of decision-making technique in supplier selection: A sysmatic review of literature.' *Expert Systems with Application,* vol 40, (2014) page 3872-3885.

[4] Fabio De Felice, Mostafa H, Deldoost, Mohsen Faizollah Antonella Patrillo. "Performance Measurement Model for the Supplier Selection Based on AHP." *International Journal of Engineering Businiess Management.* Vol 7. (2015).

[5] Pratap, Ravindra and Sharma, Shivanim, "*A Case Study of Risk Prioritization Using FMEA Method."* International Journal of Scientific and Research Publications, vol 3, 7 (2013).

[6] Chai, K., J. N. K. Liu and E.W.T. Ngai "*Application of Decision-making Technique in Supplier Selection: A systematic Review of*

Literature." Exper Systems with Applications/ vol 40 Issue 10 page: 3872 – 3885. (2013).

[7] Chan, F. T. S. and N. Kumar. "Global Supplier Development Considering Risk Factor Using Fuzzy Extende AHP-based Approach". Omega. Vol 34 Issue 4 Page : 417-431.

[8] Faris, C. W., P. J. Robinson and Y. Wind. International Buying and Creative Marketing. Boston, Allyn & Bacon. (1967).

[9] Scott, J., W. Ho, P. K. Dey and S. Talluri (2014). "A decision support system for supplier selection and order allocation in stochastic, multi-stakeholder and multi-criteria environments." International Journal of Production Economics.

[10] Cebi F, Bayraktar D. "*An integrated approach for vendor selection*". Logistics information management applied. 16:395-400. (2003).

[11] Kwang CK, Ip WH, Chan JWK. "Combining scoring method and fuzzy expert systems approach to vendor assessment: a case study". Integrated Manufacturing Systems. 13:512-519. (2012).

[12] Degraeve Z, Roodhooft F, Doveren VB. "The use of total cost of ownership for strategic procurement, a company-wide management information system." Journal of the Operational Research Society. 56:51-59. (2005)

[13] Akarte M Surendra N, Ravi B, Rangaraj N. *"Web based casting supplier evaluation using analytic hierarchy process*". Journal of the Operational Research Society. 52:511-522. (2001).

[14] Ghodsypour SH, O'Brien C. "A decision support system for vendor selection using an integrated analytic hierarchy process and linear programming." International Journal of Production Economics.56-57:199-21. (1998).

[15] Cengiz K, Ufuk C, Ziya U. "*Multi-criteria vendor selection using fuzzy AHP*". Logistics information management. 16:382-394. (2003).

[16] Pi WN, Low C. "*Supplier evaluation and selection via Taguchi's loss function and an AHP"*. International Journal of Advanced Manufacturing Technology. doi:10.1007/s00170-004-2227-z.(2005)

[17] Rao RV. "Vendor selection in a supply chain using analytic hierarchy process and genetic algorithm methods". International Journal of Services and Operations Management.; 3:355-369. (2007).

[18] Cao Q, Wang Q. *"Optimizing vendor selection in a two-stage outsourcing process*." Computers & Operations Research. 2006; doi:10.1016/j.cor. 2006.01.013.

[19] Amid A, Ghodsypour SH, O'Brien C. *Fuzzy multiobjective linear model for supplier selection in a supply chain*. International Journal of Production Economics.104:394-407. (2006).

[20] Weber, C. A., J. R. Current and W. C. Benton (1991). "*Vendor selection criteria and methods.*" European Journal of Operation Research 50: 2-18.

[21] Dickson, G. W. (1966). "*An anlysis of vendor selection systems and decisions.*" Journal of Purchasing 2(1): 12.

[22] Cheraghi, S. H., M. Dadashzadeh and M. Subramanian (2004). "*Critical success factors for supplier selection: an update*." Journal of Applied Business Research (JABR) 20(2): 18.

[23] A. F. Apriliana, R. Sarno and Y. A. Effendi, "Risk analysis of IT applications using FMEA and AHP SAW method with COBIT 5," in *International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, 2018.

[24] N. Y. Setiawan and R. Sarno, "Multi-Criteria Decision Making for Selecting Semantic Web Service Considering Variability and Complexity Trade-Off,," *Journal of Theoretical and Applied Information Technology,* pp. 316-326, 2016.

[25] A. Cahyapratama and R. Sarno, "Application of Analytic Hierarchy Process (AHP) and Simple Additive Weighting (SAW) methods in singer selection process," in *International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, 2018.

[26] R. Sarno and F. Sinaga, "Business Process Anomaly Detection using Ontology-based Process Modeling and Multi-level Class Association Rule Learning," in *International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, Bandung - Indonesia, 2015.

[27] Rezaei, J. (2016). Best-worst multi-criteria decision-making method: Some properties and a linear model. Omega, 64, 126–130.

[28] Chitsaz, N., & Azarnivand, A. (2016). Water scarcity management in arid regions based on an extended multiple criteria technique. Water Resources Management, 1–18.

[29] Salimi, N., & Rezaei, J. (2016). Measuring efficiency of university-industry Ph. D. projects using best worst method. Scientometrics, 109(3), 1911–1938.

[30] Salimi, N. (2017). Quality assessment of scientific outputs using the BWM. Scientometrics, 1–19.

[31] Salimi, N., & Rezaei, J. (2018). *Evaluating firms' R&D performance using best worst method.* Evaluation and Program Planning. 66. 147 – 155.

Strategy for Optimizing Cost and Time of After Sales using Integer Programming

Ardy Januantoro
Department of Management Technology
Faculty of Business and Management Technology
Institut Teknologi Sepuluh November
Surabaya, Indonesia
januantorro@gmail.com

Riyanarto Sarno
Department of Informatic
Faculty of Informatic
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

***Abstract*— After-sales service is a service that companies use to increase customer loyalty and trust in their products. in the handling, the company will implement two processes, namely replacement of goods or repairs. the decision to repair and replace certain items will cause problems. where the company will take optimal steps to choose to repair and replace. there are advantages and disadvantages in choosing one of these decisions. the repair process has the advantage that the cost of replacement of goods is minimal but has weaknesses in workmanship, while the replacement process has excess time but has a fairly expensive replacement cost. Continuous optimization is a way to produce results that will produce the best results that is appropriate. choices are based on variables used as decision supported. There are several methods that can be used to solve problems, one of which is Integer programming. Integer programming is a mathematical method used for optimization by maximizing or minimizing objective functions of constraints with decision variables in integers. Integer programming is chosen because it can produce replacement decisions or improvements with 0 or 1 notation. Integer programming supports decision making about replacement or repairment of returned products. The results show that the costs of the aftersales process is reduced by 12.14%, while the time is increased by 7.4%**

Keywords—optimizing; integer programming; replacement strategy;)

I. INTRODUCTION

Optimization in progress is a way to produce results that will produce the best results accordingly. Besides that, with the proper formulation, it can provide optimal decision [1, 2] There are several methods that can be used to solve the problem, one of these methods is linear programming [3]. Linear programming aims to determine and make good corrective decisions from several alternative options [4]. Linear programming can also be used in solving the problem of allocating limited resources [5]. The objective function consists of the ability to measure to minimize or maximize output while constraints are the boundaries in determining decisions. There are several types of linear programming, Integer programming is a mathematical model that allows the results of split linear programming in integer. In a previous study, Lawrie in [6] and Akkoyunlu in [7] present a linear programming model and integer programming for various types of problems and thus succeed in computing the optimal solution to the problem of the schedules of each school and university. The selection of optimum integer programming solutions is also used by P. Zwaneveld a, *, G. Verweij a, S. van Hoese in [8] to determine an optimal economic embankment investment strategy to protect against flooding. In [9] the in other cases integer programming is also used to optimize forecasts in banking cases.

In this case, the problem that must be optimized is how to minimize costs and time in the after-sales process of the company. in this study, the company used as the research subject was PT. XYZ. PT. PT. XYZ is a wireless device distributor company. The company does not have a control limit of the cost of handling the after-sales process. The aftersales process includes repair and replace. In determining repair and replace, often companies find it difficult. Repair process has advantages that are replacement cost only on broken spare parts, so the cost is cheaper. But has a long time to implement because it must use a try and check strategy. While the replace process has advantages that are a short time because it directly replaced the new item without any checks. but the cost incurred is greater than the repair process. This research will be proposed using integer programming to solve the optimization problem for choosing repair or replace. The optimization problem aims to determine which items will be replaced or repaired. Usually the result of the purpose function using integer programming is how much the maximum and minimum values. However, in this paper the mathematical method of integer programming will result in a value of 0 or 1. That is used as a decision on a good will be repaired or replaced

This paper is separated into several steps of workmanship. first is to set limit optimization using Return on Investment (ROI), second is to make optimization model and used model using integer programming method. at the end will explain the results of the research.

II. RESEARCH METHODS

A. ROI (Return of investment)

Profitability in relation to investment, using two measures of Return on Investment (ROI) and Return on Asset (ROA). ROI can measure the effectiveness and show the company ability to generate profits from the assets used [10]. In this study, ROI is used as a limitation of existing resources. The resources in this study are costs. Thus, the total costs incurred

for repairs or replacements in the aftersales process must not exceed ROI. If it exceeds ROI, the company does not benefit. The result of the calculation of ROI is usually a ratio, the ratio is used as a measure of the percentage efficiency of a company in utilizing their invested capital. Besides, with the ratio can also be labeled as the company's ability to generate payback [11]. The limits determination of this paper first determines the ROI of the company in the case study. ROI is a performance measure used to evaluate the efficiency of investment or to compare the efficiency of a number of different investments. To calculate ROI [12].

$$ROI = \frac{(total\ sales - investment)}{investment} \quad (1)$$

B. Integer programming

Integer programming is one type of linear programming method used to obtain optimal results. In addition to integer programming many methods are used for optimization. as in the case of [13] which aims to optimize the time and cost of the container port using goal programming method, goal programming can also be combined with fuzzy to optimize time and cost [14]. In this case integer programming is the most appropriate method because the optimal result is a repair or replace decision. The form of integer programming arises, because of the fact on the optimization result not all the decision variables are worth the fractional numbers as performed by linear programming. In [15], integer programming forces optimization results in the form of integers because optimization problem using 0 – 1 variable.

In integer programming, optimizing a problem will require a mathematical model. the mathematical model that will be used must represent the existing problem into a variable. In making the model on integer programming, there are three parts. There is the decision variable, constraint function and objective function

The decision variable is the variables which decide output and represent solution. For this research, the decision variable denoted by RP_n and RR_n (n = 1,2,3 ……n) where *RP* is replaced and *RR* is repaired

The constraint function is restriction or limitations on the decision, in this case the limit used ROI. The model constraint is:

$$(X_n BRP + X_n OP + X_n BRPO)x\ X_n RP + (X_n BRR + X_n OP + X_n BRRO)x\ X_n RR \leq ROI \quad (2)$$

Where:

$X_n BRP$ = cost of replace in number of item *n*
$X_n OP$ = cost of operational in number of item *n*
$X_n BRPO$ = cost of operational replace in number of item *n*
$X_n BRR$ = cost of repair in number of item *n*
$X_n BRRO$ = cost of operational repair in number of item *n*

In the model it is explained that all costs incurred in the repair or replace process are entered in one type of total cost per item, the cost is in amount with each other cost. At the end of the model it is known that the resulting model of the optimization to be generated must not exceed the set limits. The objective function is defined as the objective of making decision. The model objective function is:

$$MinZ = \sum_{n=1}^{X} RP + \sum_{n=1}^{X} RR \quad (3)$$

Where:

X : number of items.
RP : item replaces
RR : item repair

In the objective function model explained that the result of the objective function is the total of the resulting decision result

III. RESULT AND ANALYSIS

A. Set limit optimization using ROI

In integer programming limits for capacity to be optimized is important. In this case then it is known sales at the company is IDR 17,631,936.39 and the company investment is IDR 13,521,360.89 then use the following formula:

$$ROI = \frac{(17{,}631{,}936.39 - 13{,}521{,}360.89)}{13{,}521{,}360.89} = 30{,}40\% \quad (4)$$

So, the company's ROI is 30.40% from the sales that is IDR. 17,631,936.39. So, the result is IDR 5,360,215.32.

B. Optimize using Integer Programming

Before doing optimization, the first step is to determine the data that will be used in the function constraint model that has been created (Formula 2). The data used are operational cost, replacement cost and repair cost. The data are detailed in detail in order to produce a balanced and accurate results. In this research, the data used is existing data in September 2017 of PT. XYZ.

The process of data collection is done by performing data request directly to company management. Data to be taken is data related to after sales process. Such as operational cost data, data history service monitoring, data replacement cost each item, data repair each item and data sales

First, operational cost contains the cost of the operational activities on the administration process when the goods come for the service. In the operational process there are several activities as follows:

TABLE I. OPERATIONAL ACTIVITIES

No.	ID Task	Task name	Description
1	T1	Request RMA	Make a document of application service of goods from the customer
2	T2	Check status	Check the warranty status of the goods
3	T3	Info status	notify the warranty status of the goods
4	T4	Create document receipt service	Making documents for receipt of goods from customers
5	T5	Fill out the registration form	Making documents for receipt of goods from customers
6	T6	Submit the goods to the technician	Sending the goods to be service to technician

From some of the above activities, the following will explain the cost of the company for some of these activities. That cost includes computer costs, employee salaries, paper, pencils and printing costs contained in the operational process.

TABLE II. OPERATIONAL COST

ID. Task	Resources	Description	Cost
T1	Receptionist	- Computer (per/hour) - Salary (per/hour)	- IDR 3.460 - IDR 16.500
T2	Admin RMA	- Computer (per/hour) - Salary (per/hour)	- IDR 3.460 - IDR 17.500
T3	Admin RMA	- Computer (per/hour) - Computer (per/hour)	- IDR 3.460 - IDR 7.500
T4	Admin RMA	- Paper - Print - Pen	- IDR 100 - IDR 1000 - IDR 2200
T5	Admin RMA	- Computer (per/hour)	- IDR 3460
T6	Admin RMA	- Paper - Print	- IDR 200 - IDR 1000

Second, data history service monitoring is a data history about the goods from the customer entered in aftersales process. Table III is data history service monitoring and shows there are a total of 14 items that are serviced with 5 replace and 9 repair statuses.

TABLE III. ITEM ON SERVICE

No	Item name	Mac. Address	Status
1	RB 1100 AHx2 ROUTERBOARD	00:0C:02:77:12:23	Repair
2	RB 450 ROUTERBOARD	00:0C:02:88:22:22	Replace
3	RB 750 r2 (EU) hEX lite ROUTERBOARD	00:0C:02:65:66:71	Replace
4	RB 951Ui-2HnD ROUTERBOARD	00:0C:02:66:59:2D	Replace
5	RB433AH ROUTERBOARD	00:0C:02:88:4D:2A	Repair
6	RB450G ROUTERBOARD	00:0C:02:88:22:22	Repair
7	RB450G ROUTERBOARD	00:0C:02:88:22:35	Repair
8	RB450G ROUTERBOARD	00:0C:02:88:12:67	Repair
9	RB750Gr3 ROUTERBOARD	00:0C:02:65:86:1A	Replace
10	RB941-2nD-TC ROUTERBOARD	00:0C:02:55:3D:55	Repair
11	RBDynaDishG-5HacD ROUTERBOARD	00:0C:02:75:64:91	Repair
12	RBGroove-52HPn MIKROTIK	00:0C:02:55:6A:21	Repair
13	RBGrooveA-52HPn MIKROTIK	00:0C:02:55:8E:43	Repair
14	RBLHG-5nD ROUTERBOARD	00:0C:02:55:AA:7F	Repair

Third, after it is known what items will be processed in service. After it is known what items will be processed in the aftersales, next how much it costs if the item is replaced. Cost replace item as follows:

TABLE IV. REPLACEMENT COST

No	Name item	Cost
1	RB 1100 AHx2 ROUTERBOARD	IDR 3,153,692.78
2	RB 450 ROUTERBOARD	IDR 480,148.25
3	RB 750 r2 (EU) hEX lite ROUTERBOARD	IDR 400,750.32
4	RB 951Ui-2HnD ROUTERBOARD	IDR 559,953.65
5	RB433AH ROUTERBOARD	IDR 1,380,047.81
6	RB450G ROUTERBOARD	IDR 935,504.26
7	RB450G ROUTERBOARD	IDR 935,504.26
9	RB450G ROUTERBOARD	IDR 373,874.87
10	RB750Gr3 ROUTERBOARD	IDR 211,238.14
11	RB941-2nD-TC ROUTERBOARD	IDR 2,340,000,00
12	RBDynaDishG-5HacD ROUTERBOARD	IDR 549,810.00
13	RBGroove-52HPn MIKROTIK	IDR 730,560.12
14	RBGrooveA-52HPn MIKROTIK	IDR 534,772.17

Four, the cost repair used average cost of the component to be repaired. because in the process of repair costs incurred is the cost of replacing components. to calculate average cost of the component is as follows:

$$Cost\ average = \frac{Total\ of\ component\ price}{Total\ of\ item\ quantity} \quad (5)$$

$$Cost\ average = \frac{750.000}{7} = 107.142 \quad (6)$$

So, cost of each components is IDR 107.142

Cost technical replace is the cost in the load on the replace process is done. That cost includes the cost of computers and salaries of employees who handle the replace process

TABLE V. COST TECHNICAL REPLACE

No	Resources	Description of Cost	Cost
1	Admin RMA	-Computer (per/hour) - Salary (per/hour)	IDR. 3.460 IDR. 35.000

Cost technical repair is the cost in the load on the repair process. The cost is the cost to the repair technician. That cost includes the cost of computers, equipment and salaries of employees who handle the replace process. In the repair process, it takes 5 hours for the process. The process includes analysis to select the defective component, component replacement and final device testing to determine the resistance of devices that have been repaired

TABLE VI. COST TECHNICAL REPAIR

No	Resources	Description Cost	Total
1	Technical support	- Computer (per/hour) - (Salary (per/hour) * 2 technician) * 5 hours - Tools	IDR. 17.300 IDR. 175.000 IDR. 29.100

After determining data and set limit resource with ROI to be used. Next is to collect various data that had been in prepared to know result of optimal decision to replace and repair. The data that has been prepared is entered into the optimum model integer programming that has been made before (Formula 2).

Minimize:

$$MinZ = RP1 + RP2 + RP3 + +RP4 + RP5 + RP6 + RP7 + RR1 + RR2 + RR3 + RR4 + RR5 + RR6 + RR7 + RR8 + RR9 + RR10 + RR11 + RR12 + RR13 + RR14 \quad (7)$$

Subject to:

*(3153692.78 + 38,460 + 56,100) * RP1+ (480148.25+ 38,460 + 56,100) * RP2 + (559,953.65 + 38,460 + 56,100) * RP3 + (400750.32+ 38,460 + 56,100) * RP4 + (559953.65+ 38,460 + 56,100) * RP5 + (1380047.81+ 38,460 + 56,100) * RP6 + (935504.26+ 38,460 + 56,100) * RP7 + (935504.26+ 221400+ 56,100) * RP8 + (935504.26+ 221400+ 56,100) * RP9 + (373874.87+ 221400+ 56,100) * RP10 + (211238.14+ 221400+ 56,100) * RP11 + (2340000+ 221400+ 56,100) * RP12 + (549810+ 221400+ 56,100) * RP13 + (730560,12+ 221400+ 56,100) * RP14 + (534772.17+ 221400+ 56,100) * RR1 + (107.142 + 221400 + 56,100) * RR2 + (107.142 + 221400 + 56,100) * RR3 + (107.142 + 221400 + 56,100) * RR4 + (107.142 + 221400+ 56,100) * RR5+ (107.142 + 221400+ 56,100) * RR6 + (107.142 + 221400 + 56,100) * RR7 + (107.142 + 221400 + 56,100) * RR8 + (107.142 + 221400 + 56,100) * RR9 + (107.142 + 221400 + 56,100) * RR10 + (107.142 + 221400 + 56,100) * RR11 + (107.142 + 221400 + 56,100) * RR12 + (107.142 + 221400 + 56,100) * RR13 + (107.142 + 221400 + 56,100) * RR14 = 2,722,287.30 ....Constraint 1*

$RR_n + RR_m = 1$ ..... *Constraint 2* (8)

From the model can be applied directly to the software solver that will be used

Source code of integer programming

```
MODEL:
! integer programming
          optimizing cost and time;
SETS:
  PROD / X1, X2, X3, X4, X5, X6, X7/   : REPAIR, REPLACE, BRP, BRR,
BRRO, BRPO, OP;

  TUNGGAL /TUNG/ : BREPLACE, BREPAIR;
ENDSETS


 [OBJ1] MIN = @SUM( TUNGGAL : BREPLACE + BREPAIR);

 @SUM (TUNGGAL(I) : BREPLACE(I) + BREPAIR(I)) <= 2722287.30;

@FOR(PROD (PRD):
          REPAIR (PRD)+ REPLACE  (PRD) = 1);

@FOR(TUNGGAL(I) :

@SUM (PROD  (PR) :
                    (BRR  (PR)+BRRO(PR)+OP(PR)  )*  REPAIR(PR))=
BREPAIR (I);

@SUM (PROD  (PR) :
                    (BRP  (PR)+BRPO(PR)+OP(PR)  )*  REPLACE(PR))=
BREPLACE(I););


DATA:
  BRP      =
```

```
        221238.14,400750.32,559953.65,3153692.78,730560.12,534772.
17,373874.87;

BRR     =       107142, 107142, 107142, 107142, 107142, 107142,
107142;

BRRO    =
        221400,221400,221400,221400,221400,221400,221400;

BRPO    =       38460,38460,38460,38460,38460,38460,38460;

OP              =
        56100,56100,56100,56100,56100,56100,56100;
ENDDATA
END
```

Fig. 1. Source code to optimizing with integer programming

Fig.1 is the source code to optimize the aftersales process by using the integer programming method based on the previously created model

TABLE VII. RESULT OF OPTIMIZATION

No	Item Name	Mac. Address	Repair	Replace
1	RB 1100 AHx2 ROUTERBOARD	00:0C:02:77:12:23	1	0
2	RB 450 ROUTERBOARD	00:0C:02:88:22:22	1	0
3	RB 750 r2 (EU) hEX lite ROUTERBOARD	00:0C:02:65:66:71	1	0
4	RB 951Ui-2HnD ROUTERBOARD	00:0C:02:66:59:2D	1	0
5	RB433AH ROUTERBOARD	00:0C:02:88:4D:2A	1	0
6	RB450G ROUTERBOARD	00:0C:02:88:22:22	1	0
7	RB450G ROUTERBOARD	00:0C:02:88:22:35	1	0
8	RB450G ROUTERBOARD	00:0C:02:88:12:67	1	0
9	RB750Gr3 ROUTERBOARD	00:0C:02:65:86:1A	1	0
10	RB941-2nD-TC ROUTERBOARD	00:0C:02:55:3D:55	0	1
11	RBDynaDishG-5HacD ROUTERBOARD	00:0C:02:75:64:91	1	0
12	RBGroove-52HPn MIKROTIK	00:0C:02:55:6A:21	1	0
13	RBGrooveA-52HPn MIKROTIK	00:0C:02:55:8E:43	1	0
14	RBLHG-5nD ROUTERBOARD	00:0C:02:55:AA:7F	1	0

Table VII is the result of optimization on fig. 1. As for attribute information in this table is item name, mac.address and results

TABLE VIII. RESULT COMPARATION

No.	Item Name	Macc. Address	Before	After
1	RB 1100 AHx2 ROUTERBOARD	00:0C:02:77:12:23	Repair	Repair
2	RB 450 ROUTERBOARD	00:0C:02:88:22:22	Replace	Repair
3	RB 750 r2 (EU) hEX lite ROUTERBOARD	00:0C:02:65:66:71	Replace	Repair
4	RB 951Ui-2HnD ROUTERBOARD	00:0C:02:66:59:2D	Replace	Repair
5	RB433AH ROUTERBOARD	00:0C:02:88:4D:2A	Repair	Repair
6	RB450G ROUTERBOARD	00:0C:02:88:22:22	Repair	Repair
7	RB450G ROUTERBOARD	00:0C:02:88:22:35	Repair	Repair
8	RB450G ROUTERBOARD	00:0C:02:88:12:67	Repair	Repair
9	RB750Gr3 ROUTERBOARD	00:0C:02:65:86:1A	Replace	Repair
10	RB941-2nD-TC ROUTERBOARD	00:0C:02:55:3D:55	Repair	Replace
11	RBDynaDishG-5HacD ROUTERBOARD	00:0C:02:75:64:91	Repair	Repair
12	RBGroove-52HPn MIKROTIK	00:0C:02:55:6A:21	Repair	Repair
13	RBGrooveA-52HPn MIKROTIK	00:0C:02:55:8E:43	Repair	Repair
14	RBLHG-5nD ROUTERBOARD	00:0C:02:55:AA:7F	Repair	Repair

Table VIII shows the results of the comparison of existing data in table III with the data after optimization in table VII. can be seen in the item RB 450 ROUTERBOARD with mac address 00: 0C: 02: 88: 22: 22, RB 750 r2 (EU) hEX lite ROUTERBOARD with mac address 00: 0C: 02: 65: 66: 71, RB 951Ui-2HnD ROUTERBOARD with mac address 00: 0C: 02: 66: 59: 2D, RB750Gr3 ROUTERBOARD with mac address 00: 0C: 02: 65: 86: 1A the results obtained after the optimization process are from replace to repair. but in item RB941-2nD-TC ROUTERBOARD with mac address 00: 0C: 02: 55: 3D: 55 the results obtained are the opposite, that is, from repair to replace.

The results can be concluded that integer programming can produce optimization in after sales process. out of a total of 14 items in service there are total of 5 items that have changed from replace to repair or repair to replace. from the results of this study improved that the reduction in costs in the aftersales process is 12.14%. With costs before the optimization process is IDR 6,039,387.09 become IDR 5,306,144.14

IV CONCLUSION

This study shows the results of optimizing cost and time of after sales using integer programming. Four items change process from replacement to repairment, and one item changes from repairment to replacement. The change could reduce the previous cost of IDR 6,039,387.09 to IDR 5,306,144.14. Therefore, the optimization can reduce cost by 12.14%, while the time increases by 7.4%.

ACKNOWLEDGMENT

Authors would like to thank the Department of Management Technology, Faculty of Business and Management Technology on Institut Teknologi Sepuluh Nopember Surabaya for supporting this research.

REFERENCES

[1] Shoffi Izza Sabilla, Riyanarto Sarno, Yutika Amelia Effendi. Optimizing Time and Cost using Goal Programming and FMS Scheduling. 2018 International Conference on Information and Communications Technology (ICOIACT), pp. 244 - 249, DOI: 10.1109/ICOIACT.2018.8350727

[2] Nanang Yudi Setiawan, Riyanarto Sarno, Multi-Criteria Decision Making for Selecting Semantic Web Service Considering Variability and Complexity Trade-Off, Journal of Theoretical and Applied Information Technology, vol. 86, no. 2, pp. 316-326, 2016

[3] Hiller and liberman, 1982 in tarmizi. (2005). Optimasi Usaha Tani Dalam Pemanfaatan Air Irigasi Embung Leubuk Aceh besar,

[4] Izaz Ullah Khan, Norkhairul Hafiz Bajuri, Imran Abbas Jadoon. (2011). Optimal production planning for ICI Pakistan using linear programming and sensitivity analysis

[5] Kistiani, Frida. (2010). Optimasi Pendanaan Proyek Dengan Teknik Pemrograman Linier (Studi Kasus: Proyek-proyek dengan Kontrak Unit Price)

[6] N.L. Lawrie, An integer linear programming model of a school timetabling problem, The Computer Journal 12 (1969) ,307–316

[7] E.A. Akkoyunlu. (1973). A linear algorithm for computing the optimum university timetable, The Computer Journal 16 (4), 347– 350

[8] P. Zwaneveld a, G. Verweij a, S. van Hoese, (2017) Safe dike heights at minimal costs: An integer programming approach, DOI: 10.1016/j.ejor.2018.03.012

[9] Dmytro Matsypura, Ryan Thompson, Andrey L. Vasnev, Optimal selection of expert forecasts with integer programming, Omega 2017, DOI: 10.1016/j.omega.2017.06.010

[10] Sakina Ichsani, Agatha Rinta Suhardi, (2015). The effect of return on equity (ROE) and return on investment (ROI) on trading volume, DOI: 10.1016/j.sbspro.2015.11.118

[11] AL Matarneh, F. G. Financial List Analysis. (2009), An Applied and Theoretical Introduction. Dar Al Masira,

[12] Erdogmus, H., Favaro, J. and Strigel, W. Guest Editors' Introduction: Return on Investment. IEEE Software, 21(3), 18-22. 2004

[13] Shabrina Choirunnisa, Riyanarto Sarno, Abd. Charis Fauzan. Optimization of Forecasted Port Container Terminal Performance Using Goal Programming. 2018 International Conference on Information and Communications Technology (ICOIACT), pp. 332 - 336. DOI: 10.1109/ICOIACT.2018.8350719

[14] Made Agus Putra Subali, Riyanarto Sarno, Yutika Amelia Effendi. Time and Cost Optimization using Fuzzy Goal Programming. 2018 International Conference on Information and Communications Technology (ICOIACT). pp. 471- 476. DOI: 10.1109/ICOIACT.2018.8350723

[15] S. Daskalaki, T. Birbas, E. Housos. (2004). An integer programming formulation for a case study in university timetabling

Improving Business Process by Evaluating Enterprise Sustainability Indicators using Fuzzy Rule Based Classification

Lia Ninda Safitri
Departement of Management Information Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
lianindasafitri@gmail.com

Riyanarto Sarno
Departement of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Gita Intani Budiawati
Departement of Management Information Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
gitaintani2014@gmail.com

Abstract— In nowadays, business process can change rapidly. The sustainability of business process model also become consideration because of its rule in sustainability. The change of business process must rely on the evaluation based on facts. Not only use quality measurement frameworks, but also use quantity measurement frameworks. The use of event log in process mining can be as reference for evaluation of business process aligned to sustainability indicators needed. This research focuses on the classification of business processes based on sustainability indicators and business process models which have sustainability duration benchmarks. The results show that the classification of business processes based on sustainability indicators can be utilized to improve business processes.

***Keywords*— business process, sustainability indicator, classification, process mining, evaluation**

I. INTRODUCTION

Standard Operational Procedure (SOP) is fundamental for enterprises or organizations in constructing business activities. Business process is made and held correspond to SOP and selected targets.

An efficient business process makes enterprises possible to delay, change and restart the business process [6]. The change in business process itself is not necessarily possible, but there are a mechanism frameworks used to evaluate and upgrade it.

Evaluations are not only restricted on business process activity but also the concept of business process. In other word, an approach to evaluate the quality of business process is objective, quantitative and based on facts [3].

Factors contributed to produce optimal business process, directly or indirectly, can be seen from management process, operational process, and supporter of business process. Business process model can be reused as an attempt of process business management in a repository or set of business process formed [10].

The effectiveness of business process can be seen from activity relationship linkages. It is an attempt to determine the compatibility or similarity in several complex process model and to review information management flows in process model [4].

Many papers discuss the evaluation frameworks and the increasing of business process model. However, the attribution in sustainability of business process itself is slightly discussed. Mentioned by Heidari in 2014, the evaluations of business process are based on facts. The presence of log event data from business process based on true events is a reason why data from log event are used to evaluate the business process [3].

To discover the model of business process, an event log is used as starting factor of business process model analysis. Event log is a simple useful resource that helps to produce information of business process activities. [11]

Log event data in discovery will represent the business process model happened. Either the activity of the data follows SOP or not, can be seen from log event data. Log event business process model is matched to SOP business process model to know how far SOP held. Matching process is one of the techniques to identify, measure the similarity of set of business process model [2].

The similarity of business process model can be described by text, structure and behaviour similarity [10].

Some problems in measurement methods are raising when it face the worse data. Frequently, the data is not suitable with real life problems due to the human preferences so it is impossible to predict the preference in exact number. The other model such as fuzzy logic model is needed to perform such measurement because it ability to measure uncertainly such as in sustainability evaluation [12].

Business process sustainability evaluation held based on classification score of sustainability indicators. Minshu in 2014 describe how to use indicators to catch relevant aspects from business process and related phase in process management life cycle[7].

This paper will describe the way to combine the similarity evaluation of business process model results happened and the evaluation of corporation business process sustainability to do

maintenance for increasing ideal business process model correspond to addressed benchmark.

II. Previous Work

A. Similarity Graph

Research by Thada and Jaglan in 2013 shows how calculations from similarity methods from behaviour and structure[8]. This research aims to find the suitable value to collect document from website using several similarity coefficient methods such as Jaccard, Dice and Consine similarity. This research also mention that there is no best method in similarity coefficient than one methods for their research.

In other research use similarity coefficient method as use for business process method. The result is a gap of the similarity itself. There are many purpose to understand the similarity gap of business process such as Clustering, Common process finding, Scalability measurements, Flexibility measurements, Matching, Workflow finding and Fraud potential finding.

From the list above, there is no purpose that connect similarity gap to sustainability of business process model. Some questions arose for sustainability problems such as " In 5 years next, is this business process still applicable?". This question need to be answered.

B. Enterprise Sustainability

Speaking of sustainability of business process model surely rise a think of what should be done to remake or evaluate the sustainable business process model and all the thing needed to support it sustainability.

According to Doc'ecalova research in 2017 about "Evaluations of corporate sustainability indicators based on fuzzy similarity graphs", there are some indicators that can be use as references to measure sustainability [5].

In this research, 12 business processes from 12 different enterprises are compared. Each business process has a score for each subset. Then it is classified based on score resemblance of the subset. Score are differed to four levels -- very high, high, low and very low. From here, business processes that have similar sustainability indicators will be visible and classified.

Its also produce the gap results of business processes that have a good score and bad score. In addition, it give a clue of the location of the weakness of business process in indicators subset group. Here, evaluations of business process be done.

Doc'ecalova research shows how to find the sustainability indicator gap only. Otherwise, each business process has a different model. It is the thing that make business process different. Indicator subset score model also can be different. According to that, this research aims to merge that 2 reason to increase the business process score itself.

III. Proposed Method

A. Sustanbility Indicators Classification

To achieve business process sustainability goals, everything related to sustainability benchmark become concern. Business process sustainability indicators by Do' cekalova are described in Table 1.

Table 1 shows that there are 4 general indicators used namely economic, environmental, social and governance indicator. Each general indicator has supported sub-indicator.

TABLE I ENTERPRISE SUSTAINABILITY

Economic Indicators
Cash Flow (eco_1)
Return on property (eco_2)
Environmental indicators
Utilization of recycled materials and raw materials (envi_1)
Fuel utilization (envi_2)
Trash production (envi_3)
Environmental charges (envi_4)
Social indicators
Salary diversification (soc_1)
Code Ethics contravention(soc_2)
Percentage of workers protected by a collective deal (soc_3)
Occupational illness (soc_4)
Percentage of products and services for which the impact on the health and safety of customers is evaluated during their life cycle (soc5)
Spending on recognize and confirm customer satisfaction (soc_6)
Corporate governance indicators
Percentage of tactics purpose accomplished (cg_1)
Percentage of women in corporate governance (cg_2)
Donation to politicians, related institutions and political parties (cg_3)
Total of denunciation received from stakeholders (cg_4)
Sum number of distrust for disobedience with laws and regulations (cg_5)

That indicators are embedded to be scored in business process that will be classified so that each business process has score per sub-indicator.

Fuzzy rule based classification is chosen for clustering business process. Fuzzy grade membership is agreed before calculations.

The aim of clustering is to divide business process to 3 sustainability categories namely short, medium and long sustainability. Then triangle shaped grade membership is applied as Figure 1.

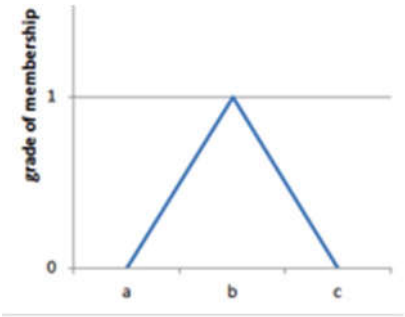


Fig. 1 Fuzzy Grade of Membership

Aligned to classification process or grouping business process for some enterprise, each business process is verbally rated (short, medium and long). Then the points of verbal score are defined to calculate fuzzy logic. Fuzzy logic dictionary is described in Table 2:

TABLE II FUZZY DICTIONARY

Verbal Score	Specification Points		
	a	b	c
Short	0	0	3
Medium	3	5	7
Long	7	8	10

Calculations of fuzzy rule based classification are performed using MATLAB 2017a.

B. *Measuring Business Process Gap using Jaccard, Dice, Consine and Wu- Palmer*

Measuring similarity gap is based on 3 approach, namely structure similarity, behaviour similarity and label text similarity. Each approach is calculated using coefficient methods which is already known such as Jaccard, Dice, Consine and [8]. Mapping of used methods for each approach is described in Table 3:

TABLE III SIMILARITY METHODS MAPPING

	Structure	Behavioral	Text
Jaccard	√	√	
Dice	√	√	
Consine	√	√	
Wu-Palmer			√

As shown in Table 3 Jaccard, Dice and Consine be used to calculate both structure similarity and behaviour similarity. According to Thada and Jaglan in 2013, there is no best calculating method of similarity coefficient so that three methods are used in this research[8].

To calculate structure and behaviour similarity, TARs (Transition Adjacency Relation set) assemblage is made from business process model compared. Tars acquired from the relation among the present activities. The relation of this pastime can also be interpreted as a sequence of execution which can occur. Within the time period referred to as the Transition adjoining members of the family [1]. In this research, existing business process model (discovery from log data) become concern. Main model (H1) will be compared to business process model that have high sustainability indicator score (H2). It can be described as:

$$H_1 = \{a_1, a_2, a_3, \dots, a_n\}$$
$$H_2 = \{b_1, b_2, b_3, \dots, b_n\}$$

Jaccard, Dice and Consine used to calculate structure and behaviour similarity are described in Table 4:

TABLE IV
SIMILARITY COEFFICIENT METHOD

Jaccard	$2\frac{X \cap Y}{X+Y}$
Dice	$\frac{X \cap Y}{X^{1/2} \times Y^{1/2}}$
Consine	$\frac{X \cap Y}{X+Y-(X \cap Y)}$

The Wu-Palmer or called WUP calculate the similarity of semantic relatedness for two word/sentences w1 and w2 in WordNet with respect to LCS (Least Common Subsummer). Similar documents in WordNet has successfully been identified by Wu-Palmer using semantic expansion. They use path length measure approach to find the similarity of two texts. [9]. The formula of WUP is :

$$Sim\ (c1, c2) = \frac{2H}{N1} + N2 + 2H$$

where :

N1 and N2 : number of is-a relation links from c1 & c2
c1 and c2 : specific common concept c
H : a number of is-a links from c to the root of the taxonomy.

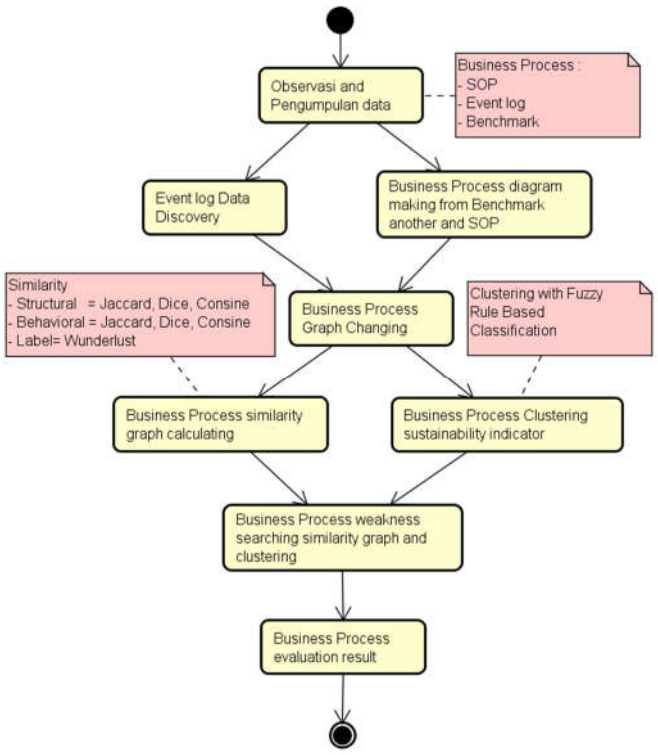


Fig. 2 Fuzzy Grade of Membership

Figure 2 explains how the flow of this research was carried out. Its need business process from event log that becomes the evaluation material, business process from another SOP and benchmark that becomes target of gap comparison. All of business process are converted to graph. After that, Its need to calculate similarity graph and clustering sustainability indicator. The results are used for evaluation and become business process improvised material.

IV. EXPERIMENT & DISCUSSION

A. *Gap of Corporate Sustainability Measurement with Fuzzy Rule Based Classification*

Research used 12 RMAs (Return Merchandise Authorization) business process of 12 enterprises. Enterprise Number 1 is a business process from event log that becomes

the evaluation material. Enterprise Number 12 is the business process from benchmark. Each business process will be rated per sub enterprise sustainability indicator. The score is given by expert judgement shown in Table 5:

TABLE V SCORE OF SUSTAINABILITY INDICATOR. ENTERPRISE NUMBER 1 CORRESPOND TO BENCHMARK AND NUMBER 12 CORRESPOND TO OBJECT OF RESEARCH

SI	Enterprise Number											
	1	2	3	4	5	6	7	8	9	10	11	12
eco1	10	5	9	9	5	8	5	6	4	8	8	6
eco2	10	7	8	9	5	8	7	5	5	8	7	6
envi1	9	3	1	7	6	7	2	7	3	4	2	3
envi2	9	2	1	7	5	7	3	6	7	5	2	2
envi3	9	1	2	7	7	7	2	7	6	3	1	2
envi4	9	0	2	7	4	7	1	8	4	7	0	1
soc1	9	2	3	8	8	8	0	2	5	6	3	1
soc2	8	2	4	8	7	8	1	3	3	4	2	2
soc3	8	1	5	7	7	7	0	1	4	5	2	3
soc4	8	3	5	7	8	7	2	0	5	3	1	2
soc5	8	2	3	9	7	7	1	2	7	4	3	2
soc6	8	1	6	8	7	8	2	3	3	5	2	1
cg1	8	2	7	8	5	7	7	7	4	7	2	2
cg2	9	5	3	7	4	7	7	4	3	3	1	3
cg3	9	3	4	9	5	7	7	3	5	4	1	2
cg4	9	4	6	8	3	7	8	6	6	3	0	1
cg5	9	7	5	8	5	6	7	4	4	5	2	3

SI = Sustainability Indicator

The enterprises will be classified using fuzzy rule based classification. Classifications are carried out in each sub-indicator namely economic, environment, social, and corporate governance.

Table 6 show the results of fuzzy rule based classification for each sub-indicator. According to table 2, score 8.3889 is classified to Long sustainability; score 5 is classified to Medium sustainability and score 1.2412 is classified to Short sustainability. Economic, Environment, Social and Corporate Governance gap from EN 1 and 12 respectively are 3.3889; 7.1477; 3.3889 and 7.1477.

TABLE VI RESULT OF SUB INDICATOR

Enterprise Number	Economic	Environment	Social	CG
1	8.3889	8.3889	8.3889	8.3889
2	5	5	5	5
3	8.3889	1.2412	5	5
4	8.3889	8.3333	5	8.3889
5	5	5	5	5
6	8.3333	5	5	1.2412
7	5	5	1.2412	5
8	5	5	5	5
9	5	5	5	5
10	8.3333	5	5	5
11	8.3333	1.2412	5	1.2412
12	5	1.2412	5	1.2412

Data in Table 6 are recalculated to obtain gap sustainability in general. The results are shown in table 7.

TABLE VII RESULT OF INDICATOR

Enterprise Number	Score	Sustainability
1	8.344	Long
2	5	Medium
3	5	Medium
4	8.344	Long
5	5	Medium
6	5	Medium
7	5	Medium
8	5	Medium
9	5	Medium
10	5	Medium
11	5	Medium
12	1.0944	Short

Can be inferred from Table 7 business process EN 1 (research object) is classified to short sustainability, while EN 12 (reference business process with highest score) is classified to Long sustainability. Gap of enterprise sustainability for EN 1 and EN 12 value is 7.2496.

B. Gap of Similarity Measurement with Jaccard, Dice, Consine and Wu-Palmer

Gap of Similarity Measurement needed business process with graph shape. Gap similarity of EN 1 and EN 12 is calculated after gap sustainability indicator is obtained. It is use to understand how far the difference of models between two business process compared. Figure 3 is the result of event log discovery from EN 12 using Neo4j. This figure constructed by 20 nodes and 26 edges. TARs assemblage EN 12 is made from Figure 3.

The result structure assemblage is represented by H1a.

H1a = start, t1, t2, t3, t4, t5, t6, t7, t8, t9, t10, t11, t12, t13, t14, t15, t16, t17, t18, t19, t20, end, A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, start-t1, t1-A, t2-B, t3-C, t4-C, t5-D, t6-E, t7-F, t8-G, t9-H, t10-I, t11-J, t12-K, t13-L, t14-M, t15-end, t16-N, t17-O, t18-P, t19-Q, A-t2, A-t4, B-t3, C-t5, D-t6, E-t7, F-t8, -, G-t9, H-t10, H-t11, H-t12, H-t16, -, J-t13, -, K-t15, K-t17, L-t14, L-t17, M-t15, N-t17, N-t18, O-t15, P-t19, Q-t17, R-t21, Q-t15

The result behavioural assemblage is represented by H1b.

H1b = AB, AC, BC, CD, DE, EF, FG, GH, HI, HJ, HK, HN, , JK, , KL, KO, LM, LO, MEnd, NO, NP, Oend, PQ, QO, REnd

The label/text assemblage is represented by H1c.

H1c = Request RMA, Check Device Status, Get Status Info, Create a service receipt, Completes RMA request form, Submit RMA request from to CSR, Device don't meet return policy criteria, Device return policy criteria, Process RMA, Issue RMA number to customer, Customer doesn't need to return device, Customer need to return device, Return Shipment, Inspects and diagnose device, Issue credit to customer.

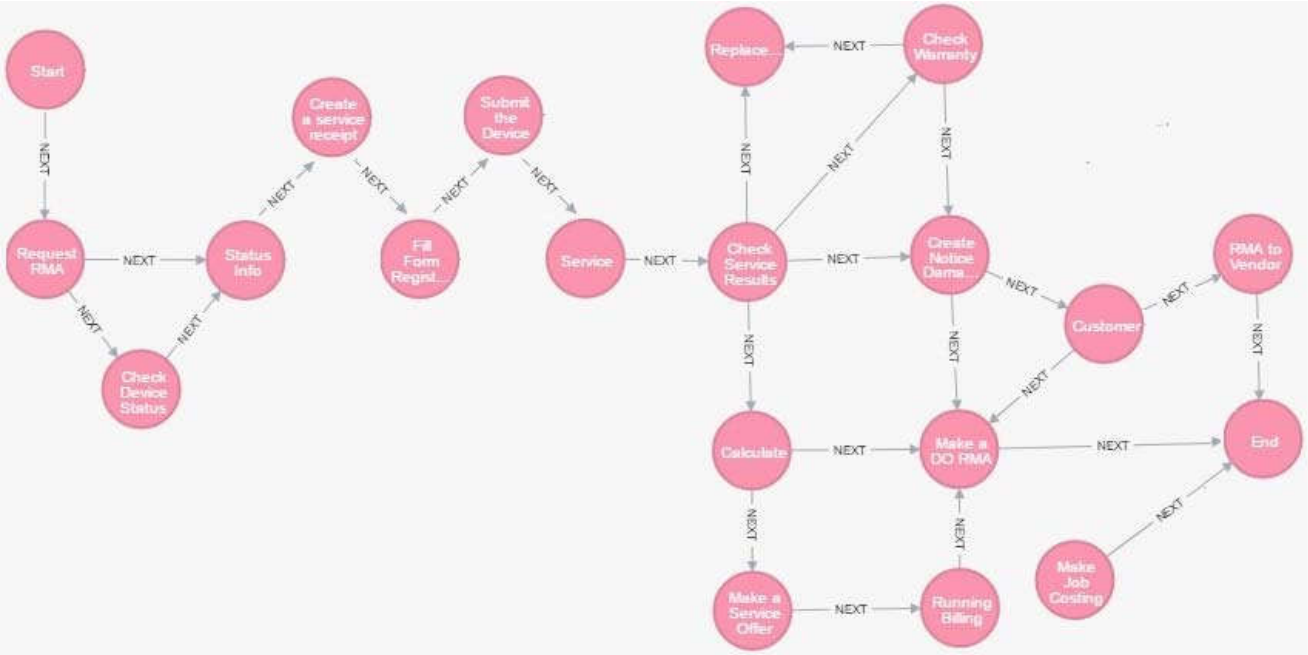


Fig. 3 Business Process of EN 12

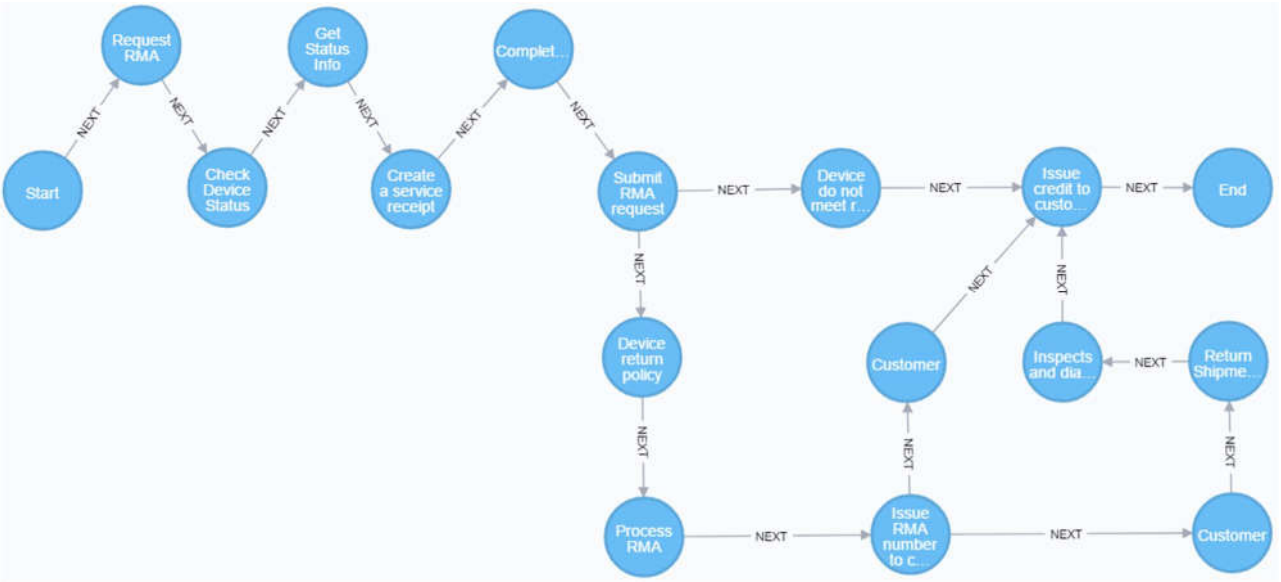


Fig. 4 Business Process of EN 1

Figure 4 is the result of modeling process business from EN 1 using Neo4j. This figure constructed by 17 nodes and 15 edges. TARs assemblage EN 1 is made from Figure 4.

The result structure assemblage is represented by H2a.

H2a = start, t1, t2, t3, t4, t5, t6, t7, t8, t9, t10, t11, t12, t13, t14, t15, end, A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, start-t1, t1-A, t2-B, t3-C, t4-D, t5-E, t6-F, t7-G, t8-end, t9-H, t10-I, t11-J, t12-K, t13-O, t14-L, t15-M, t16-N, A-t2, B-t3, C-t4, D-t5, E-t6, F-t7, F-t9, G-t8, H-t10, I-t11, J-t12, J-t14, K-t13, L-t15, M-t16, N-t13, O-t8

The result behavioural assemblage is represented by H2b.

H2b = AB, BC, CD, DE, EF, FG, FH, GEnd, HI, IJ, JK, JL, KO, LM, MN, NO, OEnd

The label/text assemblage is represented by H2c.

H2c = Request RMA, Check Device Status, Status Info, Create a service receipt, Fill Form Registration RMA, Submit the Device to the technician, Service, Check Service Results via RMA Web, Calculate the cost, Make a Service Offer, Running Billing SOP, Make a DO RMA, Check Warranty, Create Notice Damaged Device Form, Customer Info, RMA to Vendor, Replacement, Check Stock RMA, Fill the Device Request Form, Make

Job Costing, Approval Replacement, Check Sale Stock, Request Stock Expenses.

H1 and H2 assemblage is used in Jaccard with equation, Dice and Consine calculation by considering union and intersect of both. For text similarity, TARs H1c and H2c assemblage is used in Wu-Palmer using ws4jdemo.appspot.com produce average score.

The result of calculations of Structural, Behavioural and Text similarity is represented in Table 8.

TABLE VIII RESULT OF SIMILARITY GAP

	Structure	Behavioural	Text
Jaccard	0.500	0.481481	
Dice	0.583	0.45614	
Consine	0.587	0.459619	
Wu-Palmer			0.670

Jaccard, Dice and Cosine are used to calculate structure and behavioural similarity, and Wu –Palmer to calculate text/ label similarity. The average score result of that methods is 0.557 for structural, 0.466 for behavioural, and 0.67 for text label.

V. CONCLUSION

The research has shown that sustainability business process of enterprises can be predicted using enterprise sustainability indicators. The indicator score can be compared to other business process (including benchmark, as a business process references that has long sustainability) in which sustainability gap between enterprises is obtained.

It can be inferred that about 50% business process model differ from the targeted model. The business process model must be improved. The improvements are based on sustainability indicators. In addition, object business process is classified to Short sustainability so the improvement carried out by considering all of sub sustainability indicators (economic, environment, social and corporate governance). The contribution of this study has been to confirm how to calculate gap of business process and its improvement based on sustainability indicators.

VI. FUTURE WORK

Several questions still remain to be answered. Like a verbal linkage score, such as what make business process can be classified. The further research of verbal score sustainability should be done to give a better understanding about enterprise sustainability.

REFERENCES

[1] Dewi Rahmawati, Lusiana Nurul Aini, Riyanarto Sarno, Chastine Fatichah, Dwi Sunaryono. Comparison of Behavioral Similarity use TARs and Naïve Algorithm for Calculating Similarity in Business Process Model. 2017 3rd International Conference on Science in Information Technology (ICSITech), pp. 155-120. **DOI:** 10.1109/ICSITech.2017.8257095

[2] Elena Kuss and Heiner Stuckenschmidt. Automatic Classification to Matching Patterns for Process Model Matching Evaluation. 2017 CEUR Workshop Proceedings vol.1979, pp.306-319. **DOI:** http://ceur-ws.org/Vol-1979

[3] Farideh Heidari, Pericles Loucopoulos. Quality Evaluation Framework (QEF): Modeling and Evaluating Quality of Business Processes. 2014 International Journal of Accounting Information Systems 15, pp. 193-223.**DOI:** 10.1016/j.accinf.2013.09.002

[4] Klinkmüller, C., & Weber, I., Analyzing control flow information to improve the effectiveness of process model matching techniques. 2017 Decision Support Systems . pp. 6-14. **DOI:** 10.1016/j.dss.2017.06.002

[5] Marie Pavláková Docˇekalová, Karel Doubravsky´ , Mirko Dohnal, Alena Kocmanová. Evaluations of Corporate Sustainability Indicators Based on Fuzzy Similarity Graphs. 2017 Ecological Indicators Science Direct, pp. 108-114. **DOI**: 10.1016/j.ecolind.2017.02.038

[6] Mathias Weske, Business Process Management Concepts, Languages, Architectures. Berlin : Springer, 2012. **ISBN :** 978-3-642-28616-2

[7] Minsu Cho, Minseok Song, Marco Comuzzi, Sooyoung Yoo. Evaluating the effect of best practices for business process redesign: An evidence-based approach based on process mining techniques, 2014 Decision Support Systems Elsevier, pp. 92-103. **DOI:** 10.1016/j.dss.2017.10.004

[8] Muhammad Ainul Yaqin, Riyanarto Sarno, Abd. Charis Fauzan. Scalability Measurement of Business Process Model Using Business Processes Similarity and Complexity. 2017 Electrical Engineering, Computer Science and Informatics (EECSI), pp.1-5. **DOI:** 10.1109/EECSI.2017.8239129

[9] Pawan Sharma, Rashmi Tripathi, R. C. Tripathi. Finding Similar Patents through Semantic Expansion, 2016 International Conference on Computer Communication and Informatics (ICCCI), pp. 1-5. **DOI:** 10.1109/ICCCI.2016.7479982

[10] Remco Dijkman, Marlon Dumas, Boudewijn van Dongen, Reina Käärik, Jan Mendlin. Similarity of business process models : Metrics and evaluation. 2011 Information Systems Elsevier, pp. 498–516. **DOI**: 10.1016/j.is.2010.09.006

[11] Riyanarto Sarno, Yutika Amelia Effendi. Hierarchy Process Mining from Multi-source Logs. 2017 Telecommunication Computing Electronics and Control (TELKOMNIKA), pp. 1960-1975. **DOI**: 10.12928/telkomnika.v15i4.6326

[12] Uky Yudatama, Riyanarto Sarno. Evaluation Maturity Index and Risk Management for it Governance Using Fuzzy AHP and Fuzzy TOPSIS. 2015 International Seminar on Intelligent Technology and Its Applications (ISITIA), pp. 323-327. **DOI**: 10.1109/isitia.2015.7220000

IoT Communication System using Publish-Subscribe

I Made Wirawan, Irawan Dwi Wahyono, Gilang Idfi, Gradiyanto Radityo Kusumo
Department of Electrical Engineering
Universitas Negeri Malang
Malang, Indonesia
Email: made.wirawan.ft@um.ac.id, irawan.dwi.ft@um.ac.id, gilang.idfi.ft @um.ac.id,
gradiyanto.radityo.1405334@students.um.ac.id

***Abstract*— The development of the Internet of Thing (IoT) application has been very rapid in recent years. Research on IoT optimization continues to be done due to limited IoT resources. This research develops a river condition monitoring system for early warning using IoT with the Publish-Subscribe, communication model. Publisher as the sender of the river and Subscriber condition data as an application used to monitor river condition in the form of video data. The IoT Communication Model developed is the Adaptive Method. The use of the Adaptive method is due to continuous data changes and accepted by the user as monitoring data. An adaptive method is used to make efficient use of IoT resource due to big video data. Test results in this study indicate that the average efficiency of IoT resources is 34.7%.**

Keywords—IoT, monitoring, publish-subscribe.

I. INTRODUCTION

Developments in Internet computing technology are proliferating with the presence of the Internet of Things (IoT). The Internet of Things (IoT) described as a small device connected to the internet and users can control the environment or circumstances beyond the reach of users and small devices interconnected via the Internet in conducting data. Internet of Things (IoT) is a collection of small devices that work together to serve user tasks in combination [1-2]. It binds computing power to transmit data about the surrounding environment [3] [4]. The use of small devices on IoT, one of the shortcomings of the IoT is the resource of the small device.

Research on the use of IoT in monitoring system has done. Monitoring of the building using IoT has been done with low image resolution due to the consideration of the existing bandwidth so that the image quality is less useful [5]. Research on flood monitoring using wireless sensor networks by creating an IoT-based embedded system to reduce power usage and range [6-7]. Some of these studies have a disadvantage that is the excessive use of IoT resource devices that are resource processor and bandwidth.

Another study of IoT for monitoring the use of electrical power in buildings [8]. The research is only up to the accuracy of each level of voltage and current level used in the building, not doing the calculation on the IoT resource. While the use of publish-subscribe method has done on mobile devices by utilizing social media such as Facebook, Twitter and Instagram [9]. The study utilizes the internet network in performing data communication on mobile devices with social media applications. Moreover, also this research is only monitoring continuously by utilizing user or user in doing publish. The subscribed recipient is limited to only those eligible for monitoring data [9]. The resources available in the mobile device are so significant that they are not hindered in monitoring either on the publisher or subscriber side and also cannot perform continuous monitoring because of secondary publisher update data [9]. Monitoring should be done continuously to obtain details of changes in monitoring and also limited resources as they are placed in specific corners to obtain monitoring and monitoring data.

This research develops a data communication model at IoT to monitor river condition. The data communication model used is Publish-Subscribe. There are three important components in publish-subscribe that is publisher, subscriber, and broker. The publisher as the sender of river monitoring data. The subscriber as a user of river monitoring data. The publisher sends data to the subscriber based on event attraction from the subscriber. The event's interest is managed and managed by the broker. Large video data causes resource limitations of IoT devices. To overcome the limitations of resources used the Adaptive method. This Adaptive method controls the delivery of video data based on certain event interest from the subscriber. The adaptive method also arranges data sent to the subscriber is if data same with previous data hence data not sent.

II. PUBLISH-SUBSCRIBE

Publish / Subscribe is a communication mechanism or paradigm for distributed systems. Each node that is on a publish/subscribe network communicates with each other by publishing data and receiving (subscribing) data anonymously. In general, the property required by a publisher to communicate with a subscriber is the name and definition of the data. Publishers do not need any information about subscribers, and vice versa. The publish/subscribe mechanism embraces three principles. The first is space decoupling. This first principle means that the interaction between publishers and subscriber occurs anonymously. Between publishers and subscribers do not know each other's information about each other. The second is time decoupling. It means that publishers and subscribers should not be active at the same time. The publisher can publish

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

some events even though the subscribers are not currently online [10].

When the subscriber has been active again, then he will get a notification message will be a new event. The publisher can still produce events even though subscriber is not active. The third is Synchronization decoupling. It means that communication between publisher and subscriber occurs asynchronously [11]. Figure 1 shows an overview of the time and space decoupling principles provided by the publish/subscribe mechanism.

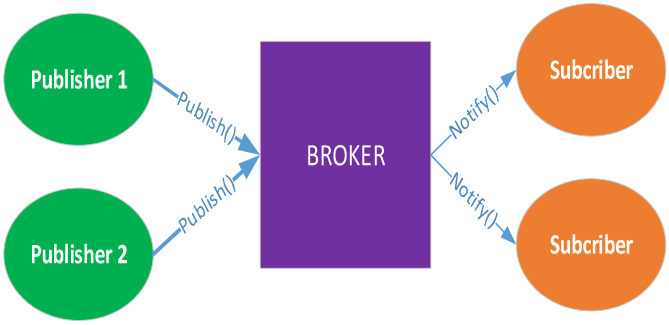


Fig. 1. The publish-subscribe mechanism

Figure 2 shows that the underlying architecture of publish/subscribe consists of 3 components. That is subscriber, broker, and publisher. The subscriber is the party that requests specific data by calling the subscribe function [10]. Subscripts submitted by the subscriber are addressed to the broker to be managed and forwarded to the data provider who publishes the data (publisher) Broker at least performs the following tasks [11]:

a. Stores the subscription

b. Accept and host events published by publishers

c. Connecting the subscriber with the publisher as a qualified data provider provided by the subscriber.

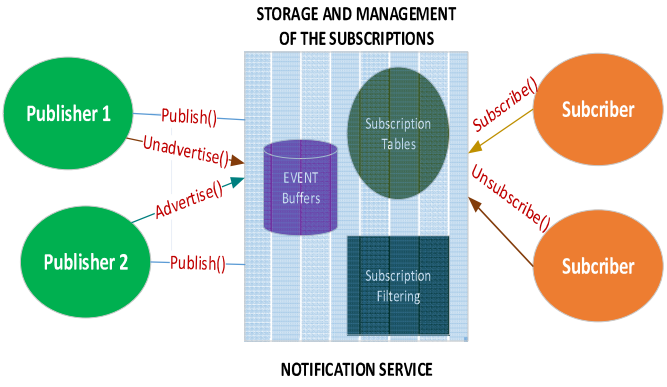


Fig. 2. The underlying architecture of publishing/subscribe

III. DESIGN SYSTEM

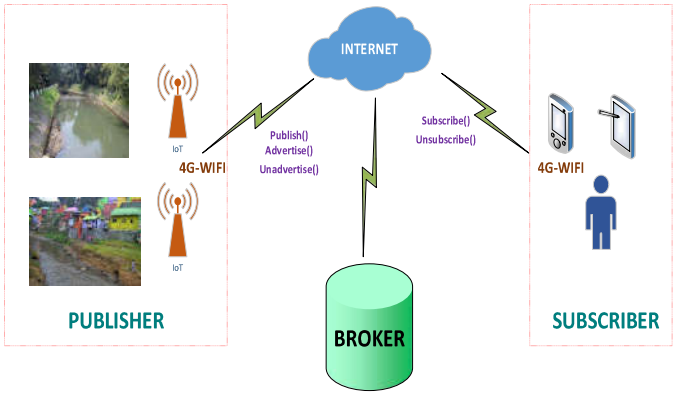


Fig. 3. The working principle of this system

This system uses the Publish-Subscribe system which consists of 3 parts, namely:

- Publisher
- Broker
- Subscriber

Publisher by utilizing IoT device that has a camera to publish the condition of the state of the road at the Universitas Negeri Malang. Publish is done by IoT device using raspberry pi 3 shown in table 1 when there is a change of image data from the previous image taken by device IoT then send to the internet and accepted by the broker with Mosca middleware. For the notification process given to the subscriber of the publisher through brokers and brokers, provide image data required by the subscriber based on data changes on the publisher. The working principle of this system described in Figure 3.

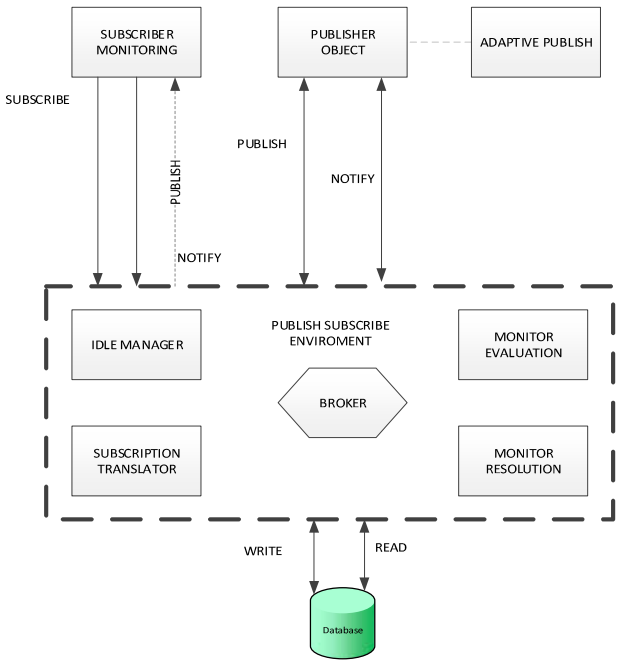


Fig. 4. The interaction between the publisher and the subscribe

TABLE I. SPECIFICATION OF RASPBERRY PI 3

Item	Type
SoC	BCM2837
CPU	64-bit 1.2GHz quad-core ARM Cortex-A53 CPU
RAM	1GB SDRAM @ 400 MHz
Storage	Micro-SD
Wireless	Integrated 802.11n wireless LAN
Ethernet	10/100 Mbps
Bluetooth	Bluetooth 4.1
Video	VideoCore IV 3D
USB 2.0	Four ports
GPIO	40
Power	2.5 Amps
OS	Linux

TABLE II. OPERATORS COMPARISON AND LOGICAL OPERATORS

Operators Comparison	Information
=	Operators equals
<>	The operator is not the same as
>	Greater than
<	Less than
>=	Larger operator equals
<=	Smaller operators equal
Operator Logika	Information
and	Logical AND operator
or	Logical operator OR

The interaction between the publisher and the subscriber can show in Figure 4. The publisher will send the event to the publish-subscribe system in the form of state information. While the subscriber sends subscripts set in the subscription language to the publish-subscribe system, the subscriber will only notify by the interest through the monitoring evaluation process. The subscription translator will translate the existing subscription. Notification to subscriber entity will handle by the broker (implemented with Mosca middleware).

In the adaptive method as shown in Figure 5, there is an idle manager that sets the publisher status to active and does the publish event to the broker or not. A subscriber (monitoring entity) to carry out the monitoring process, requires the necessary rules for communicating. The subscriber can perform monitoring under various conditions defined in the subscription translator module as shown in Table 2, which in this study is limited to location information regarding location, status and status information. Messages sent to the broker translated to then a process of evaluating the suitability of interest.

If the location information content published by the publisher matches the interest of the subscriber, the broker will notify the subscriber. The format of interest protocols is a combination of one or more conditions (merging by using logical operators), in the form of boolean expression. Example of interest format: area = 'G4' (notification of all publishers in G4 area), state = crowded and user = 'publisher1.

Characteristics of interaction in traditional publish-subscribe systems, although information from publishers is not interested in subscriber, publishers will keep publishing event information to the broker's side whenever they get new information. Each broker receives to publish information from the publisher, matching interests to each interest of the subscriber. The publishing process will be performed if there is an interest in the information from the subscriber on the adaptive method. When the publisher is declared idle by the broker, the publisher receives the event to stop the publishing process for a specified period. The publishing process will be active over if there is a match on the location information subscription list from the subscriber. While in publish-subscribe behavior in general, the matching process done for every publishing event received

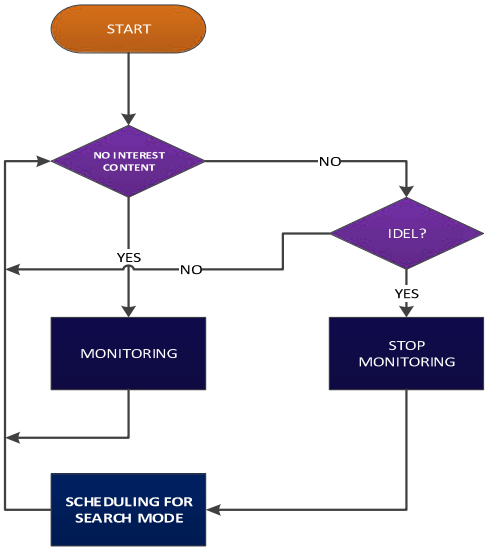


Fig. 5. Adaptive Method

IV. RESULT

Package of state/location information from the sample data, modified with additional protocols so that the event publish process will only occur for the required location information packet only (adaptive method) [12 -13]. In Table 3, we show the different publish events that occur between adaptive and non-adaptive methods in the experiment. For non-adaptive methods, the number of publishing events is not affected by the number of the subscriber. Simulation trials are conducted for 5 minutes with interval distance interval 1 second. So the number of publishing events on non-adaptive methods can be calculated by the number of publisher x 5 (minutes) x 60 (sec). In addition to the number of publish events observed, an observation of some performance parameters, including CPU server usage, server RAM usage, bandwidth usage and latency. The observation of this experiment was conducted to see the effect of the number of publishing events on the efficiency of performance. The trial did on 3G networks and Wi-Fi.

TABLE III. DATA EVENT

Publisher	Adaptive		Non-Adaptive
	5 Subscriber	10 Subscriber	
5	644	644	1500
10	712	1284	3000
15	794	1366	4500
20	864	1447	6000
25	945	1505	7500

TABLE IV. SULT TEST PERFORMANCE

Parameter	Network	Adaptive (Subscriber)		Non-adaptive (Subscriber)		Unit
		5	10	5	10	
CPU	3G	4,3829	6,7407	15,3032	18,2527	%
	Wi-Fi	4,9237	7.5088	13,9963	17,6835	
Bandwidth	3G	1805.975	2872.634	4598.5996	5246.9726	bytes/s
	Wi-Fi	1813.3332	2916.7608	4584.952	5234.6618	

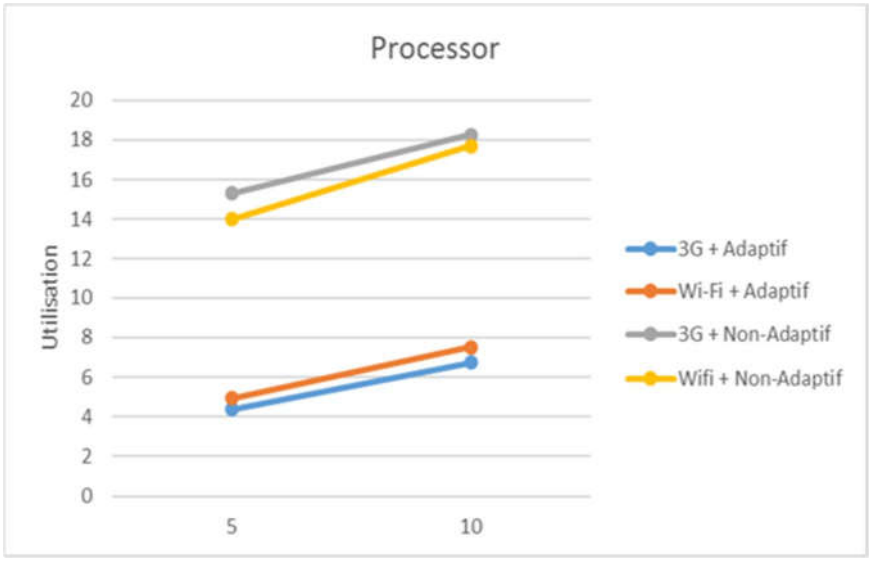


Fig. 6. The graph of efficiency CPU

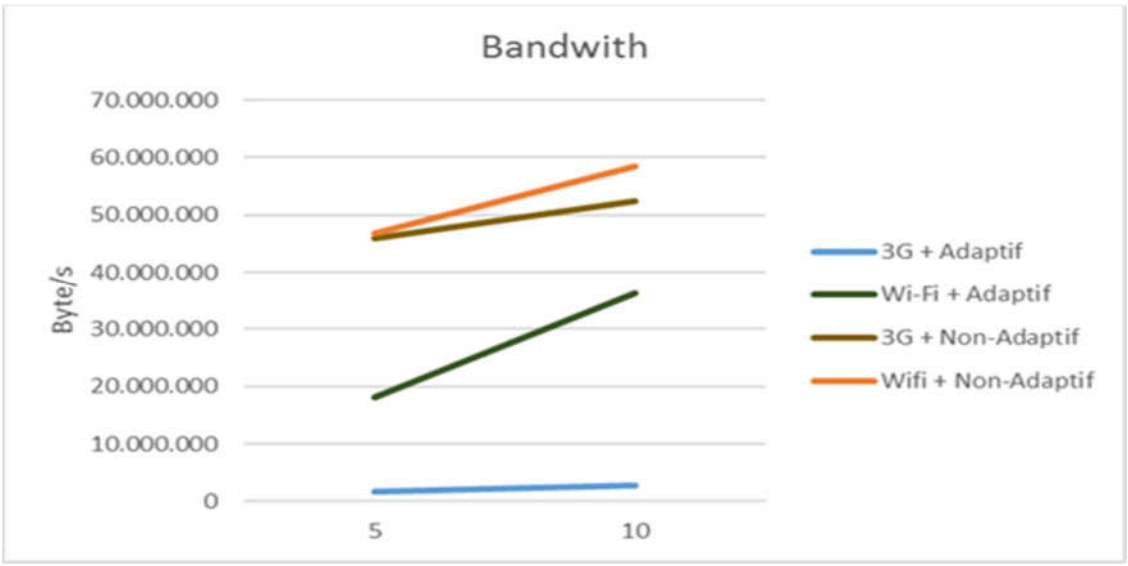


Fig. 7. The graph of efficiency Bandwith

V. DISCUSSION

The results of the performance observations shown in Table 4. The results of the experimental observations show that monitoring on IoT with the publish-subscribe model adaptively obtained for performance efficiency. For comparison of CPU usage with Adaptive model between 3G and wifi obtained efficiency equal to 10,8% by using the 3G connection. As for the use of bandwidth with the model publish-subscribe adaptive model obtained the efficiency of 91.3% using the 3G connection. For comparison of the measurement of CPU performance on the 3G connection between Adaptive and Non-Adaptive obtained by CPU efficiency equal to 66,8% for connection of publish-subscribe adaptive 3G. For comparison of bandwidth measurement on the 3G connection between Adaptive and Non-Adaptive obtained efficiency equal to 95,2% for Adaptive 3G connection. For wifi connection between Adaptive and Non-Adaptive obtained for CPU of 60,7% for Adaptive 3G usage and bandwidth 48,2% for adaptive wifi usage. The graph of efficiency shown in Figure 6 for the performance efficiency of the CPU and Figure 7 for the efficiency of the performance of the bandwidth. The indicated by bandwidth savings of 44-60% and CPU savings of 57-71%. The influenced by the existence of the idle manager module that does the active publisher checking process. In the adaptive method, the amount of location information that published influenced by the presence or absence of location information by the subscriber.

VI. CONCLUSION

This research develops data communication model at IoT to conduct road monitoring at Universitas Negeri Malang by applying adaptive publish subscribe method with attention to subscriber interest in road condition monitoring. The results of this IoT research using the adaptive publish-subscribe method in increasing the efficiency of average CPU usage by 31.76% and the average bandwidth usage is 39.4%. As the number of publishers and subscribers increases in the non-adaptive method, it will increase the usage load. The different from the adaptive method that is only influenced by the interest of information content by the subscriber.

ACKNOWLEDGMENT

This research is supported by the PNPB Grant from Universitas Negeri Malang.

REFERENCES

[1] Yin Jie; Ji Yong Pei; Li Jun; Guo Yun; Xu Wei, "Smart Home System Based on IOT Technologies," Computational and Information Sciences (ICCIS), 2013 Fifth International Conference on, pp.1789,1791, 21-23 June 2013

[2] Beligianni, F., Alamaniotis, M., Fevgas, A., Tsompanopoulou, P., Bozanis, P., & Tsoukalas, L. H. (2016). An Internet of Things architecture for preserving privacy of energy consumption.

[3] Perumal, T.; Sulaiman, M.N.; Mustapha, N.; Shahi, A.; Thinaharan, R., "Proactive architecture for Internet of Things (IoTs) management in smart homes," Consumer Electronics (GCCE), 2014 IEEE 3rd Global Conference on, pp.16,17, 7-10 Oct. 2014.

[4] Perumal, T., M.N.Sulaiman and Leong C.Y, "ECA-Based interoperability framework for intelligent building. Automation in Construction. 31, 274–280 (2013).

[5] Fadli Sirait. Sistem Monitoring Keamanan Gedung berbasis Rasberry Pi, Jurnal Teknologi Elektro, Universitas Mercu Buana, Vol. 6 No.1 Januari 2015.

[6] Mane, S. S., & Mokashi, M. K. (2015, April). Real-Time Flash-Flood Monitoring, Alerting and Forecasting System using Data Mining and wireless sensor Network. In *Communications and Signal Processing (ICCSP), 2015 International Conference on*(pp. 1881-1886). IEEE..

[7] Afzaal, H., & Zafar, N. A. (2016, October). Cloud computing based flood detection and management system using WSANs. In *Emerging Technologies (ICET), 2016 International Conference on* (pp. 1-6). IEEE.

[8] Lestari, D., Wahyono, I. D., & Fadlika, I. (2017, November). IoT based Electrical Energy Consumption Monitoring System Prototype: Case study in G4 Building Universitas Negeri Malang. In *Sustainable Information Engineering and Technology (SIET), 2017 International Conference on* (pp. 342-347). IEEE

[9] Wirawan, I. M., Herwanto, H. W., & Wahyono, I. D. (2017). Optimasi Publish-subscribe Adaptif untuk Monitoring Keamanan Selasar Gedung G4 Teknik Elektro Universitas Negeri Malang. *Jurnal Pendidikan: Teori, Penelitian, dan Pengembangan*, *2*(10), 1351-1355.

[10] Eugster, P., Felber, P., A., Guerraoui, R., *The many faces of publish/subscribe.* ACM Computing Survey., 2003

[11] Esposito, C., Cotroneo, D., & Russo, S. (2013). On reliability in publish/subscribe services. *Computer Networks*, *57*(5), 1318-1343

[12] Lopez-Vega, J. M., Povedano-Molina, J., Pardo-Castellote, G., & Lopez-Soler, J. M. (2013). A content-aware bridging service for publish/subscribe environments. *Journal of Systems and Software*, *86*(1), 108-124.

[13] Esposito. 2013. Survey on Reliability in Publish/Subscribe Service. Journal of Computer Network on (pp 1318-1343). ACM

[14] Wahyono, I. D. (2016). SERVICE DISCOVERY BERBASIS BREADTH BLOOM FILTER DI MOBILE AD-HOC NETWORK (MANET). *TEKNO*, *23*(1)..

Investigation of Ferroresonance Physical Behaviours on Three Phases Transformer with Unsymmetrical Core Leg

Mochammad Wahyudi[1,*], I Made Yulistya Negara[1], Dimas Anton Asfani[1,2], I Gusti Ngurah Satriyadi Hernanda[1], and Daniar Fahmi[1]

[1]Department of Electrical Engineering
Institut Teknologi Sepuluh Nopember (ITS)
Surabaya, Indonesia
[2]Center of Excellence for Automotive Control & System ITS
[*]mochwahyudi27@gmail.com

***Abstract*— Ferroresonance still become a confusing phenomenon since its non-linear response and very high sensitivity to parameter system and flux condition of transformer core. The existence of such greatly destructive phenomenon is basically caused by an interaction among different inductances (non-linear) with a capacitance. This paper investigated the physical characteristic of ferroresonance on three phases transformer with different dimension of core leg in which the greatest dimension was at the second phase (*S*-phase). A variation capacitance was given in order to observe the ferroresonance behaviours, that were 10 uF, 50 uF, and 100 uF. The transformer primary voltage and its total harmonics distortion (THD) were then analyzed. The results showed that the capacitance value was not linear with the ferroresonance effect level on transformer. The highest overvoltage level was seen on the test with a capacitance of 50 uF, namely about 180 %, whereas the greatest THD value was observed on the test with a capacitance of 10 uF, namely about 18 %. Moreover, the *S*-phase with the greatest core leg dimension was the most susceptible to ferroresonance as the highest overvoltage level and THD value was seen on this phase.**

Keywords—ferroresonance; unsymmetrical transformer core leg; mutual flux; total harmonics distortion; overvoltage level

I. INTRODUCTION

Ferroresonance is a power system disturbance whose appearance is still disconcerting until today since its main characteristic is highly sensitive to system parameter and initial condition [1]. More than one of steady state responses can occur for the same network parameter. In other words, its behavior is disorganized. In addition, this phenomenon is very different with linear resonance which only happens at certain values of capacitance and inductance. Consequently, the existence of ferroresonance may be not explained appropriately by a mathematics linear equation [1,2]. This is one of backgrounds leading to its appearance is unpredictable. Not only that, the presence of ferroresonance is always accompanied by extremely dangerous effects, such as overvoltage, overcurrent, and harmonics distortion in abnormal level. The presence of greatly distorted waveform may also disturb the operation of protection system [3]. Finally, no doubt if ferroresonance is categorized as one of the most dangerous phenomenon in power system.

Since the emergence of first ferroresonance term in 1920 [1], this phenomenon is defined as a non-linear oscillation in which more than one of inductance values associate with a capacitance value. The different inductance values can be observed on a saturable transformer. The inductance value in saturation area of transformer core magnetization is much lower than in normal area. The electrical faults, such as lightning strike [4-6], short circuit [5,7], switching transient [4,8-14], and unsymmetrical flux distribution [13,14] may trigger the transformer into saturation condition. On the other hands, the capacitance may come from the accumulated values of system capacitance due to line charging of transmission line, capacitor bank, capacitive voltage divider, and grading capacitance due to circuit breaker opening [1,3].

Nowadays, the ferroresonance mitigation method had widely been studied [15-24]. However, an effective mitigation system can well be designed after the physical phenomenon is understood. This can be provided by a ferroresonance test in laboratory scale in order to minimize the risk of ferroresonance not only for equipment but also for operator (human). The physical characteristics of ferroresonance and its mitigation on a single phase transformer have been investigated [25-28], whereas the test of ferroresonance initiation on a three phases transformer has not been performed yet. In fact, the transformer mostly used in field is three phases type. It means that the impact of flux mutual of three phases transformer core on ferroresonance has not be presented in the previous researches.

In this study, ferroresonance was generated on three phases transformer with unsymmetrical core leg. This modified core leg dimension would create an unsymmetrical flux distribution condition in the transformer core. The given capacitances were varied in order to know the ferroresonance characteristics, especially its sensitivity to the system parameter. The values of transformer primary voltage and total harmonics distortion (THD) for each capacitance variation were plotted into a single graph with quadrant system in order to recognize more easily the ferroresonance behavior.

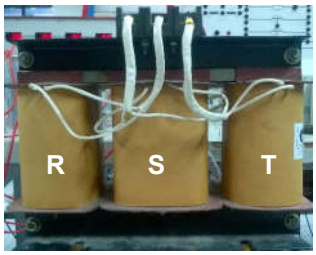


Fig. 1. The tested transformer with the greatest core leg on *S*-phase.

II. Experimental Setup of Ferroresonance on Unsymmetrical Three Phases Transformer

The feroresonance phenomenon was initiated on a three phases transformer with unsymmetrical core leg in which the dimension of *S*-phase was greater than the others as shown in Fig. 1. This would cause the unbalanced flux distribution in the core (not symmetrical). The transformer core was made of H-material type and its core shape was El type. The magnetization curve for each phase was depicted on Fig. 2. The transformer had rated power of 5 kVA and primary / secondary rated voltage of 220 V / 380 V. The windings configurations were wye on primary side and delta on secondary side.

The ferroresonance test circuit was depicted on Fig. 3. The primary windings of transformer were in series with a capacitor (V_{rated} = 450 V) and a variable power supply or a variable ac transformer (V_{output} = 0 – 380 V, f = 50 Hz, and I_{rated} = 6.5 A), whereas the secondary windings was open loop condition. There were three capacitance variations which were given in this test, namely 10 uF, 50 uF, and 100 uF. For each variation, the measurements including the root mean square (rms) value of source voltage, capacitor voltage, and transformer primary voltage and current, were conducted. In addition, the waveform of transformer primary voltage was recorded.

After the source voltage was applied to the circuit test, the value of source voltage was gradually decreased until zero in order to reduce the effect of residual flux in the transformer. The primary windings were then connected to a resistor (10 Ω) and a capacitor (40 uF). After that, the ferroresonance test was continued with other capacitance variations.

The recording of transformer primary voltage was performed by using data acquisition devices, namely Verivolt voltage sensors, NI cDAQ 9223 series, and LabView software, as seen in Fig. 4. In LabView software, the ferroresonance signals on each phase would be exported to the DIAdem software for data cutting. Moreover, the signals were transformed by using fast fourier transform (FFT) method in DIAdem. Later, the total harmonics distortion (THD) was calculated according to the obtained harmonics spectrum as equation below:

$$THD_V = \frac{\sqrt{V_2^2 + V_3^2 + V_4^2 + V_5^2 + \ldots + V_n^2}}{V_1} \qquad (1)$$

with, THD_V = total harmonics distortion for voltage signal, Vi = amplitude of voltage harmonics frequency to-i and n = a number of harmonics frequency.

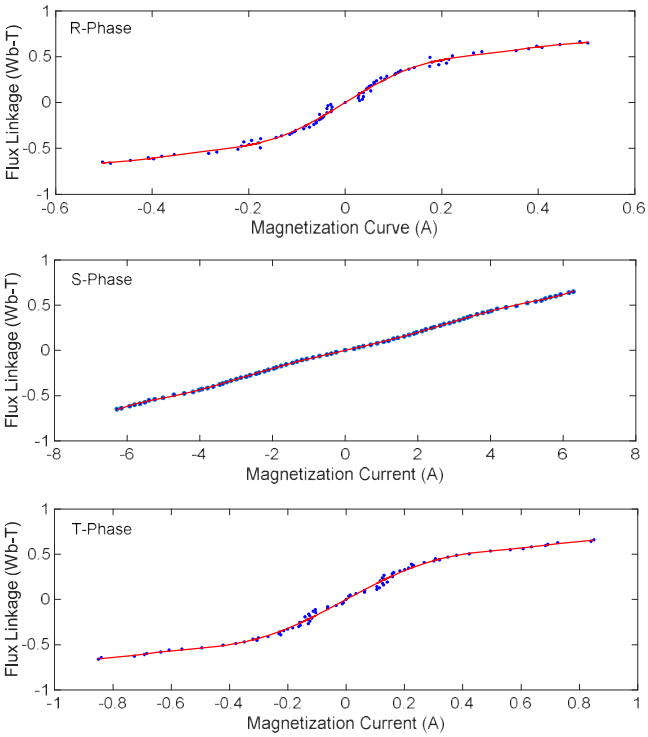


Fig. 2. Magnetization curve of three phases transformer with unsymmetrical core leg.

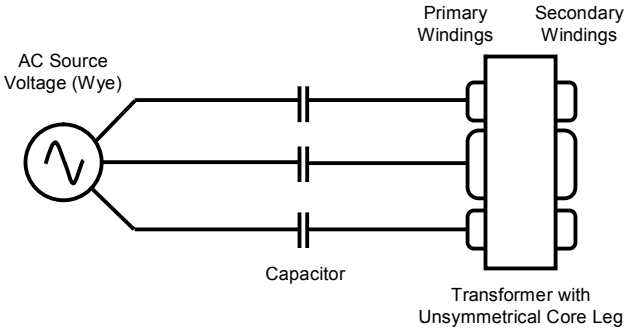


Fig. 3. Ferroresonance test circuit.

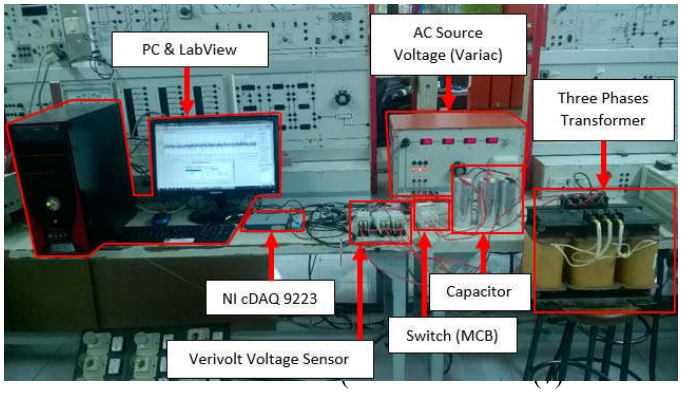


Fig. 4. Realization of ferroresonance test.

III. Results and Analysis

A. Normal Condition

The waveform of transformer primary voltage for normal condition without capacitor was shown in Fig. 5. The waveform was sinusoidal although its peak was slightly distorted due to noise from data acquisition device and source voltage harmonics. The different phase angle was 120° as the vertical line that was drawn from the peak of *S*-phase touched precisely the intersection point between *R*-phase and *T*-phase. In addition, the wave period for all voltages was 20 ms (fundamental frequency, *fs* = 50 Hz). The result of current measurement exhibited that the unsymmetrical core leg caused the current of *S*-phase to be much greater than the two other phases, namely 6.59 A at the transformer rated voltage, while the current of *R*-phase and *S*-phase was relatively the same, namely about 0.8 A.

B. Ferroresonance Condition

The feroresonance test with capacitance of 10 uF and source voltage of 200 V resulted in the transformer three phases voltage in different waveform and amplitude as depicted on Fig. 6. The distorted waveform with undervoltage condition was seen on *S*-phase, while the overvoltage condition was observed on *T*-phase. The voltage of *R*-phase tended to be in normal level, but its peak was distorted. As shown in TABLE I, the rms values of transformer primary voltage were 199.5 V for *R*-phase, 39 V for *S*-phase, and 223.4 V for *T*-phase. In addition, the difference of phase angle was no longer 120°, but about 180°.

The capacitor voltage that was higher than the source voltage was seen at the *S*-phase, namely 238 V. By regarding such voltage value, the ferroresonance tests with capacitances of 50 uF and 100 uF were conducted by applying the lower source voltage, that was about 100 V, in order to keep the capacitor voltage not to be over its rated value.

TABLE I. The Measurement Results with Ferroresonance Test Circuit: Voltage

Capacitance (uF)	Source Voltage (V)			Capacitor Voltage (V)			Transformer Voltage (V)		
	R	*S*	*T*	*R*	*S*	*T*	*R*	*S*	*T*
10	200.5	200.6	200	71.5	238	170.9	199.5	39	223.4
50	100.1	96	95.2	166.6	320	174.1	61.2	263.1	230
100	100.02	97.8	103.3	70.7	130.5	68.3	125.3	214	165.1

TABLE II. The Measurement Results with Ferroresonance Test Circuit: Current

Capacitance (uF)	Current (A)		
	R	*S*	*T*
10	0.25	0.77	0.55
50	2.55	5	2.73
100	2.18	4.05	2.12

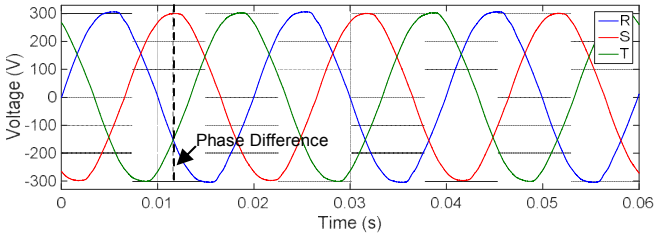


Fig. 4. Transformer voltage in normal condition with *V* source = 220 V.

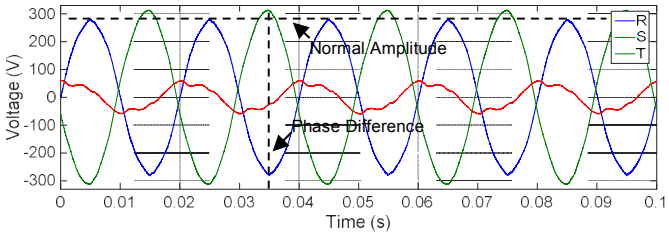


Fig. 5. Transformer voltage during ferroresonance condition with *C* = 10 uF and *V* source = 200 V.

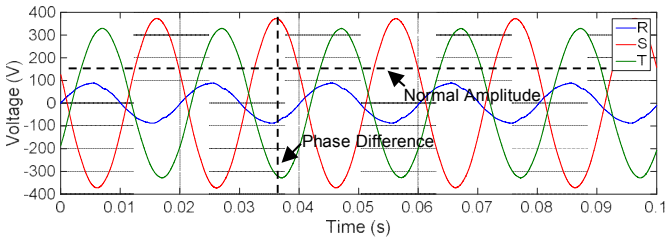


Fig. 6. Transformer three phases voltage during ferroresonance condition with *C* = 50 uF and *V* source = 100 V.

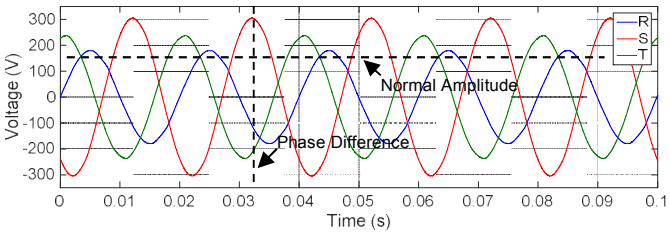


Fig. 7. Transformer three phases voltage during ferroresonance condition with *C* = 100 uF and *V* source = 100 V.

The test with 50 uF capacitance showed the overvoltage condition in significant level on *S*-phase and *T*-phase, whereas the *R*-phase was in undervoltage condition as seen in Fig. 7. The measured three phases voltages of transformer were 61.2 V for *R*-phase, 263.1 V for *S*-phase, and 230 V for *T*-phase. Besides that, the distorted waveform was observed on the *R*-phase. The shifting of phase angle also appeared in this capacitance variation.

The result of current measurement revealed that the three phases current was in abnormal value, that was 2.55 A for *R*-phase, 5 A for *S*-phase, and 2.73 A for *T*-phase as seen in TABLE II. These high currents caused the extreme increase of transformer windings temperature accompanied by the loud noise. Moreover, the capacitor voltage experienced the

significant increase compared than the previous capacitance variation. The greatest capacitor voltage was seen on the *S*-phase, 320 V, while the capacitor voltages on the two other phases were higher than the given source voltage.

The overvoltage condition on the three phases was presented by the test with 100 uF capacitance as seen in Fig. 8. The values of three phases voltage were, 125.3 V, 214 V, and 165.1 V, respectively. However, the capacitor voltage was lower than the test with 50 uF capacitance. The capacitor voltage which was greater than the source voltage was only seen on the *S*-phase. In addition, the three phases voltage waveform tended to be normal or not distorted. However, the shifting of phase angle also appeared in this capacitance variation.

The ferroresonance responses for all capacitance variations as seen in Fig. 6 - 8 were classified as fundamental mode. This was indicated by the wave period during ferroresonance condition which was 20 ms or the frequency was equal to the fundamental frequency (fs = 50 Hz).

C. Total Harmonics Distortion (THD)

The strangeness of waveform during ferroresonance could be explained by the value of total harmonics distortion (THD) as showon in TABLE III. The THD value of the test with capacitance of 10 uF was greater than the other capacitance variations. Meanwhile, the THD value of the test with capacitance of 50 uF was higher than 100 uF. This result showed that the magnitude of capacitance was opposite to the waveform distortion level.

The largest THD values were seen on the *S*-phase for C = 10 uF (18.37 %) and *R*-phase for C = 50 uF (2.68 %) as well as *R*-phase for C = 100 uF (1.17 %). The significant THD value on the *S*-phase for C = 10 uF was proved by its waveform that was worse than the other phases. Besides that, if the THD value in each phase for all capacitance variations was compared, then the largest THD was observed on the *S*-phase for C = 10 uF.

D. Mapping of Ferroresonance Behaviour

The ferroresonance behaviours were mapped into a quadrant system based on the THD value for Y-axis and the transformer voltage for X-axis as shown in Fig. 9. The IV quadrant was marked by the THD value less than 5 % (the permitted safe limit) and the transformer voltage more than its rated value (220 V). The test results with capacitance of 50 uF for two phases, namely *T*-phase (230 V) and *S*-phase (263.1 V), and capacitance of 10 uF for *T*-phase were categorized into the IV quadrant. It explained that the effect of a given capacitance value = 50 uF was very significant to create the transformer voltage more than its rated value although the given source voltage was only 100 V (lower than the given source voltage for test with capacitance of 10 uF). This condition was different with the test result with capacitance of 100 uF at the same source voltage = 100 V in which the three phases voltages was not higher than the transformer rated voltage, so that the response was classified into the III quadrant.

TABLE III. TOTAL HARMONICS DISTORTION (THD) FOR ALL CAPACITANCE VARIATIONS

THD %								
10 uF			50 uF			100 uF		
R	***S***	***T***	***R***	***S***	***T***	***R***	***S***	***T***
3.18	18.37	2.26	2.68	1.03	1.06	1.17	0.61	0.81

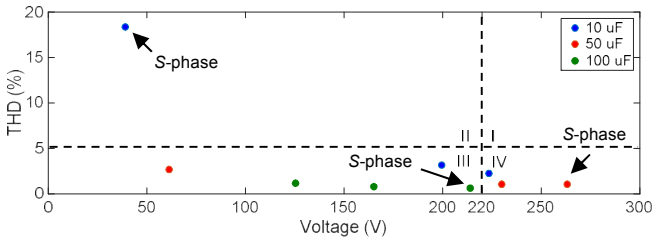


Fig. 9. Categorization of ferroresonance behaviours based on transformer rms three phases voltage (V) and THD (%).

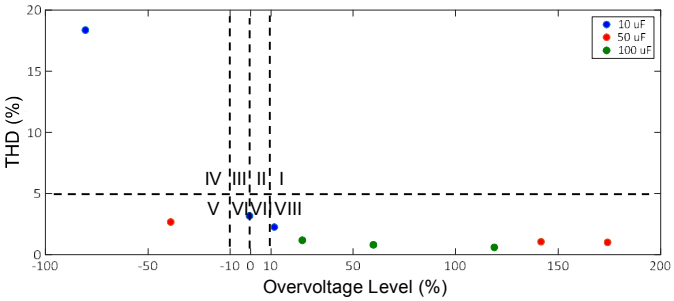


Fig. 10. Categorization of ferroresonance behaviours based on transformer overvoltage level (%) and THD (%).

On the other hand, the voltage of *T*-phase on the test with capacitance of 10 uF that was 223.4 V in the I quadrant was quite reasonable since the given source voltage was about 220 V. However, the using of capacitance of 10 uF was so significant on the distortion level of waveform as the THD value on the *S*-phase was greater than 5 %, namely 18.37 %.

The measured transformer voltage during the test with C = 100 uF which was below its rated value was actually in abnormal overvoltage level (more than 10 % or above the permitted safe limit) as seen in Fig. 10. This mapping of ferroresonance behaviour in Fig. 10 involved the overvoltage percentage from the given source voltage. In addition, the map showed that the greater capacitance did not mean the ferroresonance response getting worse. It could be observed by the test result with C = 10 uF giving the greatest THD value, so that the response was in the IV quadrant. Moreover, the test result with capacitance of 50 uF led to the higher overvoltage than capacitance of 100 uF as the two greatest overvoltage percentage (points) in the VIII quadrant was the test result with capacitance of 50 uF.

IV. CONCLUSIONS

The given capacitance variation was able to generate the ferroresonance response with different effect. In general, most of the test results resulted in overvoltage level above 10 %. The

impacts of ferroresonance on transformer, such as total harmonics distortion (THD) and abnormal undervoltage or overvoltage, were more clearly observed on the *S*-phase with the greatest dimension of core leg. The greater capacitance did not mean the impact of ferroresonance getting worse according to the values of THD and overvoltage. The test results with capacitance of 50 uF actually provided the higher overvoltage compared to the test with capacitance of 100 uF, but the overvoltage during test with capacitance of 100 uF presented the greater value than capacitance of 10 uF. On the other hand, the capacitance variations that created the THD value from the highest to the lowest were 10 uF, 50 uF, and 100 uF, respectively. In addition, a phase difference during ferroresonance condition was not 120^{o} as the normal condition. Those non-linearity phenomena proved the existence of flux mutual affecting the ferroresonance response.

REFERENCES

[1] Ferraci, Ferroresonance, Group Schneider, Cahier, no. 190, pp. 1-28, 1998

[2] Emin Z., Al Zahawi B.A.T., Auckland D.W., Tong Y.K., "Ferroresonance in electromagnetic voltage transformer: a study based on nonlinear dynamics", IEE Proc-Gener. Transm. Distrib., vol 144, no. 4, 1997.

[3] Price E., A tutorial on ferroresonance, 40th Annual Western Protective Relay Conference, October, 2013.

[4] Hernanda I.G.N.S., Negara I.M.Y., Soeprijanto A., Novandi. Wahyudi M., "Ferroreonance characteristics due to lightning strike on capacitive voltage transformer", International Review on Modelling and Simulations, vol. 9, no. 4, 2016.

[5] Sinuraya J.S.P., Negara I.M.Y., Hernanda IG.N.S., Analysis of capacitance effect on ferroresonance of power system (in Indonesian), ST Final Project, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, 2016.

[6] Bakar A.H.A., "Analysis of lightning-caused ferroresonance in capacitor voltage transformer (CVT)", Electrical Power and Energy System, vol. 33, pp. 1536-1541, 2011.

[7] Piñeros J.F., Vélez J.A., Rodríguez D., "Ferroresonance in a 115 kV network due to a single line fault. International Conference on Power Systems Transients", Cavtat, Croatia, pp. 15-18, 2015.

[8] Emin Z., Al Zahawi B.A.T., Auckland D.W., Tong Y.K., "Ferroresonance in electromagnetic voltage transformer: a study based on nonlinear dynamics", IEE Proc-Gener. Transm. Distrib., vol 144, no. 4, 1997.

[9] Sanaye-Pasand M., Aghazadeh R., Mohseni H., "Ferroresonance occurrence during energization of capacitive voltage substations", IEEE Power Engineering Society General Meeting, 2003.

[10] Akinci T.C., Ekren N., Seker S., Yildirim S., "Continuous wavelet transform for ferroresonance phenomena in electric power systems", Electrical Power and Energy Systems, vol. 44, no. 403–409, 2013.

[11] Graovac M., Iravani R., Wang X., McTaggart R.D., "Fast ferroresonance suppression of coupling capacitor voltage transformers", IEEE Transactions On Power Delivery, vol. 18, no. 1, 2003.

[12] Huang S.J., Hsieh C.H., "Relation analysis for ferroresonance of bus potential transformer and circuit breaker grading capacitance", Electrical Power and Energy System, vol. 51, 2013.

[13] Moses P.S., Masoum M.A.S., "Modelling ferroresonance in asymmetric three-phase power transformer", Australasian Universities Power Engineering Conference, 2009.

[14] Hernanda I.G.N.S., Negara I.M.Y., Soeprijanto A., Asfani D.A., Fahmi D., Andarini K., Wahyudi M., "Detection of ferroresonance on asymmetric three phase transformer due to capacitance variation". AUN/SEED-Net Regional Conference on Electrical and Electronics Engineering, 2017.

[15] Ajaei F.B., Sanaye-Pasand M., Rezaei-Zare A., Iravani R., "Analysis and suppression of the coupling capacitor voltage transformer ferroresonance phenomenon". IEEE Transactions on Power Delivery. October, vol. 24, no. 4, 2009.

[16] Khan S.A., Bakar A.H.A., Rahim N.A., Tan ChiaKwang., "Analysis of ferroresonance suppression and transient response performances for various ferroresonance suppression circuits in capacitive voltage transformers". 3rd IET International Conference on Clean Energy and Technology, 2014.

[17] Shahabi S., Gholami A., Mirzaei M., Taheri S., "Investigation of performance of ferroresonance suppressing circuits in coupling capacitor voltage transformers", 4th IEEE Conference on Industrial Electronics and Applications, 2009.

[18] Badrkhani Ajaei F., Sanaye-Pasand M., Rezaei-Zare A., Iravani R., "Analysis and suppression of the coupling capacitor voltage transformer ferroresonance phenomenon". IEEE Transactions on Power Delivery, vol. 24, pp. 1969, 2009.

[19] Sakamuri J., Yesuraj D.J., "Modeling and simulation of capacitor voltage transformer transients using PSCAD/EMTDC", IEEE Trondheim Power Tech., pp. 1-8, 2011.

[20] Rezaei S., "Impact of transmission line and plant outage on ferroresonance in AC transmission system and new suppression method", 13th IET International Conference on AC and DC Power Transmission, 2017.

[21] Ryzhkova Y.N., Tsyruk S.A., "Ferroresonance suppression in distribution networks", 2nd International Conference on Industrial Engineering, Applications and Manufacturing, 2016.

[22] Rezaei S., "Mitigation of ferroresonance by facts in electrical network", International Journal on Electrical Engineering and Informatics, vol. 9, no. 1, 2017.

[23] Yang M., Sima W., Duan P., Zou M., Peng D., Yang Q., Duan Q., "Electromagnetic transient study on flexible control processes of ferroresonance", International Journal of Electrical Power & Energy Systems, vol. 93, pp. 194-203, 2017.

[24] Yang M., Sima W., Chen L., Duan P., Sun P., Yuan T., "Suppressing ferroresonance in potential transformers using a model-free active-resistance controller", International Journal of Electrical Power & Energy Systems, vol. 95, pp. 384-393, 2018.

[25] Roy, M., dan Roy, C.K, "Experiments on ferroresonance at various line conditions and its damping", IEEE, 2008.

[26] Milicevic, K., Vinko, D. dan Vulin, D., "Experimental investigation of impact of remnant flux on the ferroresonance initiation". Electrical Power and Energy Systems, Vol. 61, Hal. 346–354, 2014.

[27] Sima, W.X., Yang, M., Yang, Q., Yuan, T. dan Zou, M., "Experiment on a novel method for fundamental ferroresonance suppression", World Scientific Publishing Company, Modern Physics Letters B, Vol. 28, No. 5 1450035, 2014.

[28] Milicevic, K., dan Emin, Z., "Initiation of characteristic ferroresonance states based on flux reflection model", IEEE Transactions On Circuits And Systems—II: Express Briefs, Vol. 60, No. 1., 2013.

Autonomous Controller for MetOcean Monitoring Station

Rizki Adi Nugroho[1*], Andrianshah Priyadi[2], Anan Fauzi[1], Taufan Wiguna[1]
[1]Laboratory for Marine Survey Technology, [2]National Laboratory for Energy Conversion Technology
Agency for the Assessment and Application of Technology
*rizki.adi@bppt.go.id

Abstract— Marine environment is an important part of coastal community life. The environmental changes that occurred along with the development of activity and infrastructure development. It is necessary to conduct environmental monitoring by making autonomous controller for MetOcean monitoring station based on single board computer with AWS as meteorology sensor and ADCP sensor. The system works in gathering data from AWS and ADCP sensor. ADCP and AWS will be serialized to text data and transmitted to server through gsm modem. The output is observed weather parameters include temperature, pressure, humidity, speed, wind direction, solar radiation, and rainfall. While the physical parameters of the ocean waters are tidal, speed, direction of surface currents, surface temperature, salinity, wave height period. For the first time, this system was installed and tested in a port area.

Keywords—MetOcean; Meteorology; Oceanography, Monitoring Station; Marine, Single Board Computer

I. INTRODUCTION

MetOcean is an observation of meteorology and oceanography parameters. Meteorological parameters consist of wind speed, wind direction, air temperature, humidity, rainfall, air pressure, dewpoint and oceanographic parameters consist of sea level elevation, surface current, and waves. All parameters are used by oceanographer to create a model and analyze its impact.

To help oceanographer get real-time, fast, and continuous MetOcean data from a remote monitoring station, the autonomous control system is needed. The system is not only gather both data, meteorological data from automatic weather station-rain gauge (AWS-RG) sensor on land and oceanographic data from Acoustic Doppler Current Profiler (ADCP) sensor in the sea, but also send it to the user. This process can be done by acquisition controller.

The controllers for data collection and monitoring system have been used in various fields, such as subway system [1], medical [2], water quality monitoring [3], transportation tracking [4], and road pricing [5]. These controllers collect data only from one sensor. Similar autonomous system for real-time monitoring weather has been applied to show how information get from instruments are effectively combined in urban area, New York, America [6], Indonesia [7], Brazil, which is this system aims at establishing weather station with different sensors, using central box to synchronize its data and equipped with weather sensors and GPS [8]. These system mention before has not touched the maritime section, as is this metocean controller, we try to make controller that can gather atmhospheric and maritime data. Focus on this paper will explain the design of a controller that can collect data from different instruments, especially marine sensor, software design, and output data**.**

The MetOcean Monitoring Station has 4 main components to run its system. They are an oceanographic sensor named Acoustic Doppler Current Profiler (ADCP), a meteorological sensor named Automatic Weather Station and Rain Gauge (AWS-RG), power management to manage devices electrical power, and data management. This system has been installed in port area at Penajam Paser Utara (PPU), East Kalimantan.

II. SYSTEM AND CONCEPT DESIGN

The process to get and gather data from sensors are described here along with autonomous controller system design.

A. Weather Measurement Sensor

This is a typical sensor for weather measurement at certain height position using MetPak RG from Gill. This sensor is equipped with components with ultrasound for parameters, such as wind speed (m/s), wind direction (0^0 till 360^0), temperature (Celsius or Fahrenheit), humidity (%RH), Rain Gauge (mm), air barometric pressure (hPa), DewPoint (Celsius or Fahrenheit). Considering weather openness location which is there is no barrier to building, hill, or tree that can make refract measurement of wind direction and wind speed. Other reasons are security and easy access at this time (for a while) this sensor was installed on rooftop PT. ASDP at the Penajam ferry port, district Penajam Paser Utara (PPU), East Kalimantan ($1^0$14’35.70’’LS and $116^0$46’36.65’’BT)

B. Current Surface and Waves Measurement Sensor

Survey tool which is used to measure current surface and sea waves is based on acoustic Doppler. Side-looking/horizontal type Acoustic Doppler Current Profiler (H-ADCP) SL-500 from sontek, USA. Acoustic Sensor consists of two horizontal beams for surface current measurement, one vertical beam to measure waves and water pressure and be equipped with a data connection and power supply (SPF-1500) along 200 m that connect this sensor with power supply management. Measurable parameters among others current speed and direction at a distance of 120 m in front of the attached sensor, water level or tidal (m), significant wave height (m), wave period (m). This sensor attached to the pillar located under the first dock at ferry port Penajam (Position: 1014'30.49''LS and 116046'36.89''BT). The location of the dock about 150 m from ASDP office Penajam where power management is attached on. It was tethered at the pillar with the help of a wooden support pole held by two circular steel clamps. The depth position of the sensor is about 4.8 m from the bottom of the where the lower position is 1.2 m from the low tide level at that location.

C. Power Management

Power management has a function to keep electrical load stable in its voltage division and also battery usage settings. It uses two system (12 V and 24 V system) to manage its load. For its own input, its uses panel solar connected to the power management. The output is DC voltage (direct voltage) 12 V and 24 V, then for the AC load, it uses inverter from 12 VDC to 220 VAC. Power Management component consist of solar panels, battery, and regulator component.

.

D. Data Managing Device

In acquisition system and traffic data is done by Data Controller and GPRS Modem.

- Data Controller

  The Data controller was built from Beaglebone and technical component. It is one of opensource single board computing hardware that has linux Angstrom ARM default operating system. Equipped with memory ability 512 mb DDR3, Ethernet on board that enable beaglebone black can connect with the internet without adding Ethernet modul anymore. The most extraordinary that beaglebone black has so more multifunction pin in right and left side of the board.
- General Package Radio Service Modem (GPRS)
  Modem is an abbreviation of modulator and demodulator. It has a function to deliver data in the form of carrier signal in order to send to the user via certain media. The process is called modulation. Demodulator is a functioning process to get back the data which is sent by the sender. In this process the data will be separate from high frequency, and analog data will be converted back to the digital signal so it can be read by a computer. The data was sent from acquisition system then sent to the office continuously (programmable), under read down station coordination.

E. System Design

The autonomous controller system consist of Beagle bone placed in ADSP office that acts a data controller. As seen on fig 1. It connected to, solar panel (power management), aws-rg sensor in rooftop ADSP building and ADCP sensor in a port and communicating via data cable to data controller and power management. Beaglebone dependent to them to get acquisition data before it compile into one and sent it to server data. Then do further analysis of the data before the show to the public.

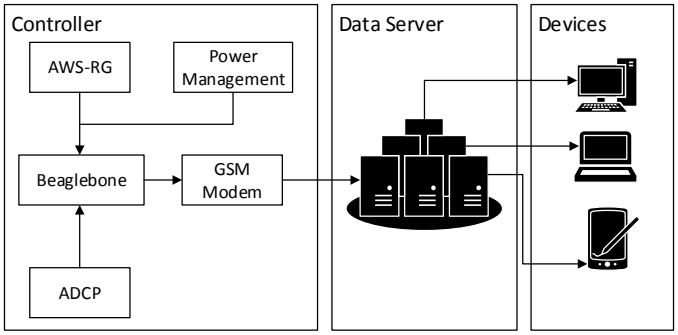


Fig 1. System Design

Beaglebone connected to these sensors and communicating directly via serial data cable. USB to serial 232 was used to communicate between AWS-RG, GSM Modem, and power management. Because ADCP position over 100 m, USB to serial 485 was used to keep good data when taken by controller. The controller collects data from adcp sampling ie date, water level, water temperature, deviation standard, power level, speed, direction, waves, internal flow and 10 ADCP profiling number in the following sequence data ie the 1st is cell number, the 2nd is velocity 1, the 3rd is velocity 2, the 4th is standard error of velocity 1, the 5th is standard error of velocity 2, the 6th is signal strength beam 1, 7th is signal strength beam 2, besides that data AWS RG ie wind direction, wind speed, wind pressure, humidity, temperature.

This system works concurrently where every 15 minutes the controller get data sampling from ADCP and AWS-RG in text format and save autnomatically in external memory card installed in single board computer. After all the required data is collected as needed, it is the task of the gsm modem to transmit the data every 60 minutes to the server data in Jakarta. So there is 4 sampling data AWS-RG and 4 sampling data ADCP plus 10 current profile in 60 minutes.

F. Software Design

The autonomous controller software consist of several modul split into 4 parts main parts: adcp handler, aws handler, gsm handler, file handler.

- Adcp handler

  Works to do all related parts sea wave and tidal sensor installed below sea level. Some function do how to read adcp sensor, time synchronization, average interval, sampling interval, signal handler adcp, get data sampling, open device sensor, close device sensor, send query sensor adcp.
- Aws handler

  Aws handler do all related parts about windspeed (m/s), wind direction ($0\text{-}360^0$), temperature (in celcius or fahrenheit degree), humidity (% RH), rainfall (mm), barometric pressure. Some function do how to poll aws sensor, sampling aws_sensor, wait prompt, signal handler aws sensor, open device sensor, send query aws sensor.
- Gsm handler

  GSM handler do all related parts about data transmission, signal quality, communication via sms from operator and vice versa. Some function do how to open gsm mode, wait prompt, signal handler I/O, send command, send file, reupload file, download file, write data ftp, network registration, signal quality, get error, check, send text, read text, time synchronization, delete message.
- File Handler

  File handler do all related parts with read, write file as sampling data storage was done by AWS and ADCP sensor. Some function do how to write user, read user, write initial, read initial, write status, read status, write aws data, write adcp data, read aws data, read adcp data.

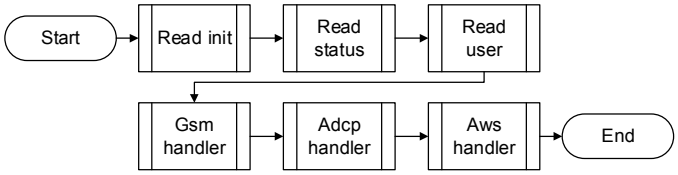


Fig 2. Flowchart metocean autonomous controller

```
int main( ){
   int dum_main;
   pthread_attr_t attr_main;
   pthread_t thrd_adcp_handler;
   pthread_t thrd_gsm_handler;
   signal(SIGINT, finish);
   STOP = FALSE;
   dum_main = system("clear");
   //inisialisasi parameter, status dan user
   readInit(&parameter);
   readStatus(&status);
   readUser(user_system);

   fprintf(stdout,"Program start\n");fflush(stdout);

   pthread_mutex_init(&count_mutex, NULL);
   pthread_attr_init(&attr_main);
   pthread_attr_setdetachstate(&attr_main,
   PTHREAD_CREATE_JOINABLE);
   //proses gsm handler dijalankan
   pthread_create(&thrd_gsm_handler,&attr_main,(void*)gprs_handler,
   NULL);
   //menunggu sinkronisasi waktu dengan sistem
   sleep(10);
```

```
 //proses adcp handler dijalankan
pthread_create(&thrd_adcp_handler,&attr_main,(void*)adcp_handler,N
ULL);
//adcp_handler();
//proses aws handler dijalankan
aws_handler();

pthread_cancel(thrd_adcp_handler);
pthread_cancel(thrd_gsm_handler);
pthread_join(thrd_adcp_handler,NULL);
pthread_join(thrd_gsm_handler,NULL);
pthread_attr_destroy(&attr_main);
pthread_mutex_destroy(&count_mutex);
pthread_exit (NULL   );
return 0;
}
```

Fig 3. Main program controller

III. SYSTEM TEST

The system test is divided into two parts ie, dry testing that ensures the readiness of each sensor in recording data is as expected without a single data void. Wet testing was done for the whole system is ready to work automatically where the sensors are installed in accordance with the place and data transmition to the server.

```
root@beaglebone:~/programming/tes_adcp# minicom -D /dev/ttyO4 -b
9600,n,8,1
Welcome to minicom 2.6.1
OPTIONS: I18n
Compiled on Jul  3 2013, 21:16:04.
Port /dev/ttyO4
Press CTRL-A Z for help on special keys
>start deploy
Checking Setup Parameters...
OK
2015 10 16 13 38 16    2.4    0.1   2.134   0.9   0.3  10.0 102  95 157   0
136.9   0.4   0.8  0.0  0.0  0.0  30.95      1.669    0.003  11.4   1.5  113.5
27  23  16    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0    0.0
```

Fig 3. minicom dry test controller software

minicom is text based, serial communication program design for linux. we are trying to know the output of ADCP sensor and this means that ADCP is ready to deploy. AWS connected via RS-232 serial cable in dev/ttyO2 . Table 1 show measured parameters from adcp.

Table 1. Test Value ADCP sensor

No	Description	unit	Value
1	Year Month Day Hour Minute Second		2015 10 16 13 38 16
2	WaterVel1/X/E WaterVel2/Y/N	(cm/s)	2.4 0.1
3	Water level	(m)	2.134
4	VelStDev1/X/E VelStDev2/Y/N VelStDev3/Z/U	(cm/s)	0.9 0.3 10.0
5	SNR1 SNR2 SNR3	(dB)	102 95 157

6	SignalAmp1 SignalAmp2 SignalAmp3	(counts)	0 136.9 0.4
7	Noise1 Noise2 Noise3	(counts)	0.8 0.0 0.0
8	Ice Detection		0.0
9	Heading Pitch Roll	(deg)	30.95 1.669 0.003
10	Standard deviation of the Heading Pitch Roll	(deg)	11.4 1.5 113.5
11	Mean Tempr Mean Press	(degC) (dBar)	27 23
12	StDevPress	(dBar)	16
13	Power level (battery voltage)	(Volts)	0.0
14	CellBegin CellEnd	(m)	0.0 0.0
15	Speed	(cm/s)	0.0
16	Direction	(deg)	0.0
17	Wave amplitude for 10 Period Bands	(cm)	0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 5.3401
18	Significant Wave Height	(cm)	222.5079
19	Peak Period	(s)	1618658.306
20	Internal Flow-Area	(m3)	14.8651
21	Internal Flow-Flow	(m3/s)	0

This ADCP equipped with a feature to allows recording a multiple-cell velocity profile. It was setting till 10 velocity profile to know. When multi-cell profiler is enabled, the adcp will report up to eleven velocity values — one for the standard integrated cell and up to ten cells from the velocity profile. Table 2 show measured ADCP profiling velocity in beam column 1.

Table 2. Test Value profiling 1 adcp

No	Description	Unit	Value
1	Cell Number	(cm/s)	1
2	Velocity 1/x	(cm/s)	4.5
3	Velocity 2/y	(cm/s)	-1.3
4	Standard error of Velocity 1/x	(cm/s)	2.9
5	Standard error of Velocity 2/y	(cm/s)	1.3
6	Signal Strength (Beam 1)	(counts)	153
7	Signal Strength (Beam 2)	(counts)	156



Fig 4. Gill Met View Desktop dry test software

Using gill instrument metview desktop software to ensure that AWS-RG is ready to combine with the autonomous system, checking wind direction, wind speed, pressure, humidity, temperature, dew point and rainfall. Few setting like changing baudrate, setting reporting mode in continuous or polled, but in this system we use polled mode because while keep on sampling, it only provide data according to demand of controller at that time.

Table 3. Test value aws

No	Description	Unit	Value
1	Wind Direction	degree	349^0
2	Wind Speed	m/s	0.07
3	Pressure	Mbar	1002.3
4	Humidity	%	48%
5	Temperature	degC	25.50
6	Dew Point	degC	13.7^0
7	RainFall	mm	0

The testing device sensors is embedded on beaglebone black single board computing. AWS connected via RS-232 serial cable in dev/ttyO2. Using vim text linux, navigate to directory that test aws sensor.Inside folder aws modul testing there is an executable program to test aws sensor. The result of wet testing like in Fig 4. The order data indicates aws data sampled.

```
root@beaglebone:~/programming/tes_aws#  ./test  /dev/ttyO2./test
/dev/ttyO2

Q 154.000000 0.070000 1009.700012 84.300003 28.700001 25.799999
0.000000 13.200000
Q 206.000000 0.480000 1009.700012 84.300003 28.700001 25.799999
0.000000 13.200000
Q 271.000000 0.710000 1009.700012 84.300003 28.700001 25.799999
0.000000 13.100000
Q 345.000000 1.220000 1009.700012 84.300003 28.700001 25.799999
0.000000 13.400000
```

Fig 5. Wet Test AWS Sensor

Q shows node letter, wind direction, wind speed, pressure, humidity, temperature, dewpoint, rainfall, supply.

In wet test, only few parameter such as datetime, watervel1, watervel2, waterlevel, heading, pitch, roll, mean temperature, mean pressure, stddev pressure, significant wave height, peak period, internal flow area, flow and 10 profiling adcp. Parameters are activated in accordance with the needs of the researcher who will perform further data processing

```
root@beaglebone:~/programming/tes_adcp# ./adcp /dev/ttyO4
./adcp /dev/ttyO4
Kirim : ENTER
2016/06/16 17:01:42.00
GAGAL: timeout
Kirim : ENTER
2016/06/16 17:01:49.00
GAGAL: timeout
Kirim : +++
Kirim : ENTER
>ENTER
Kirim : DATE 2016/06/16
DATE 2016/06/16
OK
Sinkronisasi Tanggal Berhasil
Kirim : DATE
DATE
2016/06/16
>
Kirim : TIME 17:02:07
TIME 17:02:07
OK
Sinkronisasi Waktu Berhasil
Kirim : TIME
TIME
17:02:08
>
setting sampling interval [2016-06-16;17:05:00]
Kirim : ENTER
ENTER
Kirim : si 900
si 900
OK
Sampling Interval Berhasil di Set
Kirim : TIME
TIME
17:05:02
setting average interval [2016-06-16;17:05:00]
Kirim : ENTER
ENTER
Kirim : ai 300
ai 300
OK
Average Interval Berhasil di Set
Kirim : TIME
TIME
17:05:08
>
Kirim : ssu
ssu
OK
SSU Berhasil di Set
```

```
2016 06 16 17 11 28    8.6   -3.3 -1.000   0.2   0.0   0.0 78 112   0
 0 169.7   0.0 -0.8  0.1  0.0  0.0 30.90    4.355   00
 1   3.3   -2.2  0.5   0.1 114127
 2   5.8   -2.3  0.5   0.1 80  123
 3  10.8   -3.7  0.5   0.2  48  109
 4  18.9   -4.1  1.2   0.5  36  89
 5  26.4   -1.5  2.7   1.2  32  77
 6  27.3    1.8  3.6   1.7  31  77
 7  21.9    0.6  4.6   2.1  30  65
 8  30.7    9.3  4.9   2.3  30  52
 9  20.1    9.2  5.2   2.4  30  47
10   7.3   11.1  5.0   2.3  30  46
```

Fig 6. Wet Test ADCP Sensors

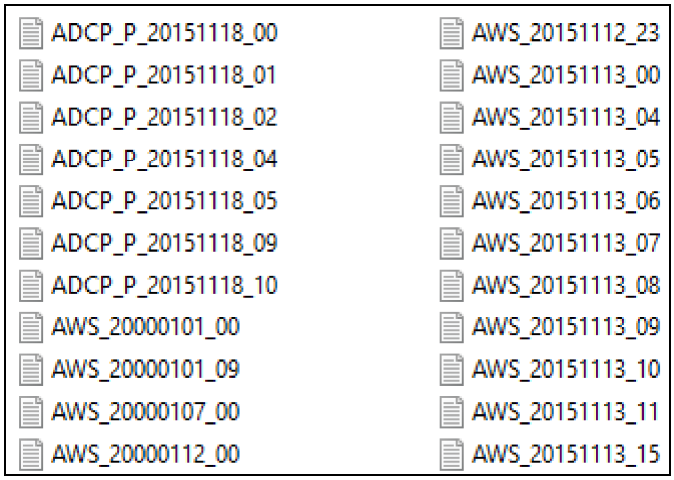


Fig 7. Data stored inside internal storage of Beaglebone

The data that stored on storage beaglebone have its own format according to the device and the sampling date. Beside that, the data also sent to the data server in Jakarta using ftp with the help of gsm modem. Next data will be parsed with parser system, to view on web application for further data utilization.

IV. CONCLUSION

The autonomous control system that are able to record all aws, adcp sampling every 15 minutes and send it to server every 60 minutes has been presented. The prototype autonomous controller for metocean monitoring station works successfully until the data saved in internal storage of beaglebone and send it to data center via ftp. This show that is possible to use beaglebone for autonomous metocean monitoring station and using internet data packet to communicate. A few improvements have to be made including repairing status, parameter, reboot system via remote using sms.

Acknowledgment

This is 2015 in maritime technopark activities from Pengkajian dan Penerapan Teknologi Survei Kelautan (PP Teksurla) Programme. The author would like to thank to the installation technical team of autonomous controller for

metocean, ASDP office, laboratory for marine survey technology, the major of Penajam Paser Utara District.

References

[1] Alessandro Carbonari, Massimo Vaccarini, Mikko Valta and Maddalena Nurchis (June 8th 2016). Wireless Real-Time Monitoring System for the Implementation of Intelligent Control in Subways, Real-time Systems Kuodi Jian, IntechOpen, DOI: 10.5772/62679

[2] Huang Jinshan, "Design of real time monitoring system for dangerous transportation based on DSP", Procedia Engineering 174 (2017) 1323-1329, 2017 [13th Global Congress on Manufacturing and Management, GCMM 2016]

[3] C. Leibold, M. Wahalla, C. Blume, H. Blume, and M. Wilhelmi, "A real-time monitoring system controller for medical tissue engineering bioreactors," Biomedical Circuits and System Conference (BioCAS), 2015.

[4] Daudi S. Simbeye, "Shi Feng Yang.Water Quality Monitoring and Control for Aquaculture Based on Wireless Sensor Networks," Jurnal of Network 9(4):840-849, April 2014

[5] Lundy Orlando Tjandra, Saptadi Nugroho, Darmawan Utomo," Electronic Road Pricing System Prototipe", 2016 [International Seminar on Application for Technology of Information and Communication]

[6] F. Lo Conti, D. Pumo, A. Incontrera, A. Francipane, L.V noto, G. L Loggia. A Weather Monitoring System For The Study of Precipitation Fields, Weather, and Climate in an Urban Area. 2014[11th International Conference on Hydroinformatics]. New York, USA

[7] D. Krisnandar, "Perancangan dan analisa output rangkaian signal conditioning analog melalui mikrokontroller atmega8535 untuk stasiun cuaca," INKOM, vol. 5, no. 1, pp. 22–28, 2011

[8] Ahmed El-Deeb, Aislan Gomide Foina (October, 2008). PEWEMOS – Pervasive Weather Monitoring System, Conference: IEEE Third International Conference on Pervasive Computing and Applications, DOI: 10.1109/ICPCA.2008.4783562

Optimization Service Discovery in Wireless Balloon Network

Irawan Dwi Wahyono, Irham Fadlika, A.N. Afandi, M. Rodhi Faiz, Sunarsih, Dwi Prihanto
Department of Electrical Engineering
Universitas Negeri Malang
Malang, Indonesia
Email : irawan.dwi.ft@um.ac.id, irham.fadlika.ft@um.ac.id, an.afandi@um.ac.id, Mohamad.rodhi.ft @um.ac.id
sunarsih.1405334@students.um.ac.id, dwi.prihanto.ft @um.ac.id

***Abstract*—** Due to some limitations and constraints in computation, power, bandwidth, high mobility on every node and determination of central node coordination, development of service optimization research discovery on Mobile Devices at MANET (Mobile Ad-hoc) and (WBN) Wireless Balloon Network always done to get the optimal result. In this research, optimization in Wireless Balloon Network (WBN) using Breath Bloom filter proposed. The service descriptor packet for service request and advertisement in layer network is optimized using Breadth Bloom Filter to reduce the packet size. The distribution of packet on the network layer performed by utilizing routing protocol AODV. The method of developing the optimization in WBN (Wireless Balloon Network) using Breath Bloom filter is called WBN-P. Data analysis and test results of the simulation shows the success of WBN-P in overcoming constraints on WBN, which are the reduction of bandwidth of 18.6% and 30.9% delay.

Keywords—Wireless Balloon Network, Discovery, AODV, bandwidth, delay

I. INTRODUCTION

The Wireless Network Balloon Network (WBN) is almost identical to the Mobile Ad-hoc network of mobile users with different functions and uses, various types of services, things, applications, sensors, and many resources that work together [1-2]. Limitations, constraints, and complexity in wireless ballon networks are same communication mode in MANET network especially in Ad-hoc communication mode [3-4]. Many ways of solving problems in Ad-hoc communication mode, one way to solve a user provides all the elements of service. In this case by automatically sharing and accessing regardless of location and their ownership of the service function. It happens mainly in the discovery of each node in ad-hoc communication mode [5-6].

From the previous study, there are several Service Discovery methods, one of which is Konark [7]. Konark is implemented on the application layer protocol side regardless of the packet carrier for distribution at the routing protocol layer. One disadvantage of the Konark method and framework is the too long data packets in describing the service descriptor. The longer the distributed data packets can cause the overhead of the node. The overhead causes in MANET can also attribute to the selection of routing protocol methods for the distribution of data packets as carriers. Routing protocols are of two kinds: reactive and proactive. Therefore a method that can make the length of the data packet for the service descriptor in the application layer protocol becomes short and can also determine the carrier's preferred carrier routing at the network protocol layer[8-9]. To connect between protocols to interact with both the application layer protocol and the network layer is also called cross-layer protocol [8-10].

In this research, the solution is developed in optimization various application of wireless balloon network. However, this study aimed explicitly at solving limited bandwidth problems and high delay on wireless balloon networks. This research develops a new protocol optimization for service discovery at WBN. In this WBN, develop cross-over application layer protocols and network layers that work the following way: classification of service descriptors in tree form and defined using Breath Bloom (BBf) filters on the application layer. Service deployment is done by piggybacking on service information using network protocol routing, i.e., routing protocol AODV.

II. BREATH BLOOM FILTER

BBf (Breath Bloom filters) are representative of SD (Service Descriptors) arrange $S = \{x_1, x_2, \ldots, x_n\}$ where n is the whole service. The definition of BBf is as follows: v is declared as an array of bit m. Every bits $\{1, .. m\}$ are first arrange to 0. Filters use k own value, hash function $h_1 the\ , h_2, \ldots, h_k$, with scope $\{1, .. m\}$ for every hash SD x to array v. For every SD, the hash output of $h_i(x)$ represents the position of the array in $v, v\ [h_i(x)]$ which is declared to 1 for every hash functions $i = 1, 2, \ldots, k$. Each setting in v is assigned a value of 1 for several times [11].

To check if there is a service z in the BBf, determine whether all $h_i(z)$ is arrange to 1. If one case, then the k service is available. If all $h_i(x)$ is not 1, this service is not part of the BBf, BBf may generate FP (False Positive) if the BBf indicates that it is a SD (Service Descriptor) $k \in S$. The opportunity of obtaining a FP lookup can be measurable using the probability calculus [11].

The Breath Bloom filter (BBf) is equal to the Bloom filter which is a different array only in the hash function model in the placement of the k value where the k value is placed randomly by utilizing MD5. As for BBf which includes all services in detail in the form of arrays or pointers. BBf declaration in the form of array pointers [11] show in Fig.1 for service printer declarations.

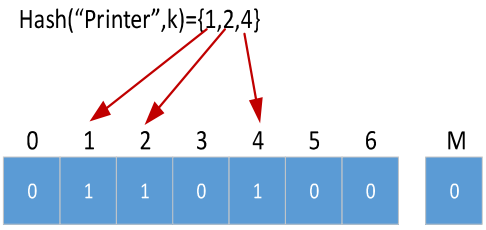


Fig. 1. BBf in array pointer

A. False Positive (FP) Calculation

The FP variable : m expressed the length of the BBf bit, n n expressed in the amount of service descriptions in the BBf, and k expressed in the amount of hash functions used, the possibility of FP given by equation 1 [12].

$$P_{\text{fp}} = (1 - e^{-\frac{kn}{m}})^k \qquad (1).$$

Note that the amount of services is the pure value that can vary while the app is running. It is, therefore, necessary to have a thorough knowledge of the app object and to initialization k and m variable to reduce the possibility of FP queries [12].

There is two mode to subtract the possibility of FP [12-13]:

1. Alter the amount of k hash functions.
2. Add the dimensions of the BBf itself,

B. OPT (Optimal) Function of Hash Function

The best worth of k can be determine by pickup the generation from equation 1 and then ian nvention that the best worth of hash, k_{opt}, m is functions for the BBf width and n is the number of SD so presented by equation 2 [13]:

$$k = \frac{m}{n} = \ln 2 \Rightarrow k_{\text{opt}} = [k] \qquad (2).$$

When developing an optimization of service discovery protocol based on BBf is essential to choose the best worth of hash functions.

III. DESIGN SYSTEM

In this study, the protocol design system architecture that will implement in NS-2 is called WBN-P. The components contained in WBN-P shown in Fig. 2, where WBN-P interacts via AODV as a packet carrier to be simulated on NS-2. AODV serves to broadcast the packet by utilizing MPR flooding techniques. In WBN-P there are four main components namely:

1. *Breath Bloom Filter or BBf* serves as a conversion of all data services
2. *Service Function* serves as addressing service requests and advertisement services from external nodes and local nodes.
3. *Message Creator* works to create a new message before it broadcast in the network of MANET *Service Advertisement (SA)* and *Service Request (SR)*
4. *Repositories* as data service storage

In the Repository component 3 blocks of the function are:

1. *Local Service* or local node service also contains a list of services owned by local nodes.
2. *External Service* contains a list of services on external/foreign nodes
3. *The requested service* contains the requested service from the node itself

The working principle of WBN-P show Fig. 4 is as follows:

Condition- 1: If a node wants to discover service S which is described as text, node make service request utilizing Breath Bloom filter (BBf) by converting service description into a small packet in the form of BBf. The form of s the ervice request is $B(SR)$. External Service repositories performs service queries on $B(SR)$, if the service is not found then the node creates a new message request again containing all service requests that are in the Request service repositories into single service requests using BBf to $B(SR)$, then the node sends the service request via WBN network by utilizing AODV routing protocol.

Condition 2: If the node receives $B(SR)$ from external node then it is queryed with local repositories containing the service list in the local node. If service S is found, the node responds promptly by doing *Service Advertisment*. It is created using BBf which consists of all services contained in Local Services repositories into $B(SA)$ and then sent to the network using AODV routing protocol with MPR flooding technique.

Condition 3: If a node wants to do *advertisement* of service, service description S is immediately added to Local Service Repository, then *service advertisement* is created and consists of all services contained in Local Service repository by using BBf to $B(SA)$ then sent to WBN network. When the node receives service advertisement $B(SA)$ from the network, the node immediately adds, updates and queries to External Service repositories.

A. WBN-P in the Breath Bloom filter (BBf)

In the use of BBf which is a hash function in which each hash function is described to map model such as SD (service descriptor) to form pseudorandom number in range $1 \ldots m$. The results on different hash functions k have to stand alone. One mode to develop a hash function in BBf is to utilize the modulo hash function. Another way is to utilize a cryptographic like MD5 [13].

The MD5 calculation process is not expected to be safe for some cryptographic applications but keep the desired properties. It is the basis of the BBf hash function in this study. The MD5 algorithm has a deterministic and irregular character and also has excellent collision endurance. The use of MD5 in this study as open source code for plentiful coding programming and its implementation is comparatively quick. Because of its grade, the possibility of FP can be solved using [13].

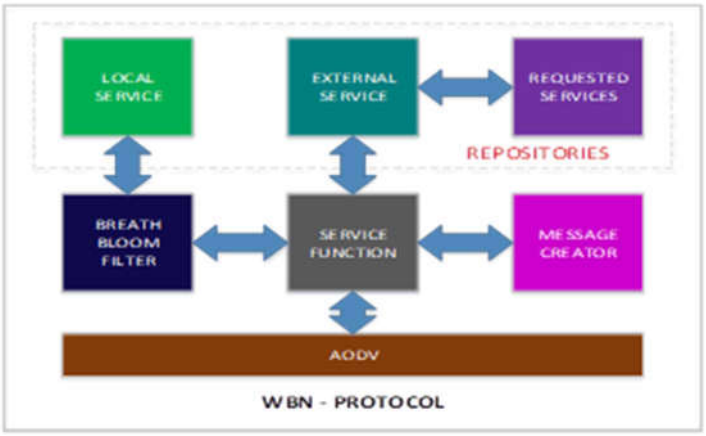


Fig. 2. Design System

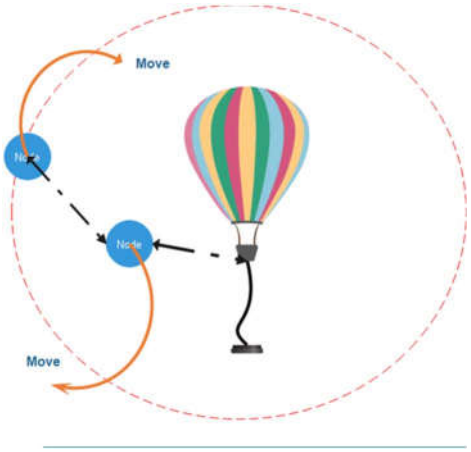


Fig. 3. Node Movement Scenario on WBN

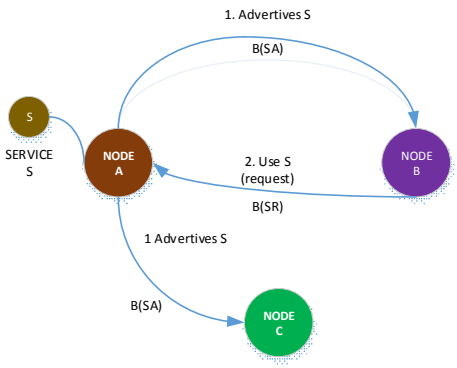


Fig. 4. The working principle of WBN-P

The cryptographic hash function particularly MD5 tries to inhibit/slow down, so computing is more delayed than general aim hash functions. However, the MD5 process only runs on *service advertising* and *service requests* and not when performing matching services as matching by the filter itself. Furthermore, In one MD5 operation, it is necessary to generate inputs for all random hash functions k. BBf declaration on the system as explained in Fig. 5. As for the declaration of service discovery using BBf is shown in Fig. 6.

```
///Pointer pada  local Breath Bloom filter (BBF)
bitset<BLOOMFILTER_SIZE>* ::getLocalBloomFilter()
{
        return &_localBloomFilter;
}
```

Fig. 5. BBf declaration

```
string ::getmd5Hash(string* service)
{
       MD5 context;
       context.update((unsigned char*)service-
>c_str(),service->size());
       context.finalize();
       char* digest = context.hex_digest();

       string line = digest;
       return line;
}

unsigned long int ::ConvertHash2Int(string hash)
{
       std::stringstream stream;
       stream << hash;
       unsigned long int n;
       stream >> std::hex >> n;
       return n;
}
```

Fig. 6. Discovery declaration

The design of WBN-P Service Discovery (MSD) uses MD5 in the following way as shown in the flowchart of Fig. 7: the hash k function, which is the BBf, is constructed from the k group of each output bit r of a 128-bit hash on operation MD5. Every sub-bit *set* of MD5 output can be function as input for its own function. Every function k is set to one bit in filter v.

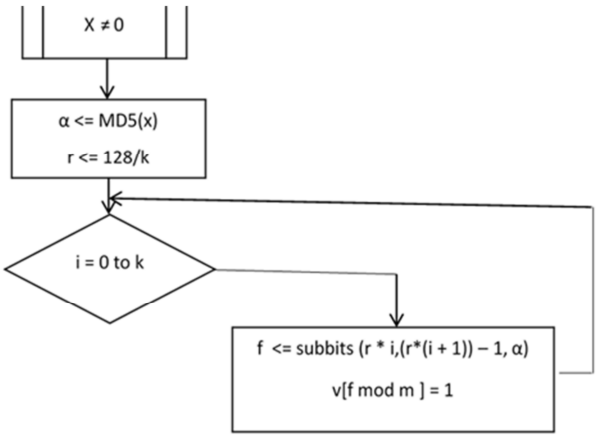


Fig. 7. Flowchart determination of k and v values

IV. TESTING

In testing WBN-P, two parameters are [14]:.

1. *Delay* evaluation (time consumption) to perform service discovery on MANET network.
2. *Bandwidth* is the number of bytes on the node generated during the service discovery process and tested utilize dissimilar network topologies.

To get the above parameters by taking advantage of two scenarios, i.e., static. Static scenarios are efficiently designed and repeated and also flexibly measured when different features of the protocols with scenarios of movement of the nodes as shown in Fig. 2 is moving freely. The measurements are done using Network Simulator NS-2.35 with default parameters used in Table 1[15]. Setting in AODV, we use default setting [16].

TABLE I. DEFAULT SETTING FOR TESTING ON NS-2.35

Parameter	Value
Simulator	NS-2.35
OS	Ubuntu 12.04
TR	100m
MAC Address	802.11
RFM	Two Ray Ground
Mov	Random Waypoint
RP	AODV
AODV setting	Default

A. Bandwidth Measurement Scenario

One various static topology device used for size bandwidth. The topology consists of nodes oriented at the square of nodes $\{4, 9, 16 \ldots 64\}$ [14]. Fig. 8 shows the test settings for 16 nodes. Every topology has two services, sited at nodes 0 and 1. At the request of the Service is done randomly by another node with the 15's interval during application running from the 1500s. For every static topology, 40 programs executed, and 95% intervals are estimated and presented in numbers. WBN-P configured with no caching to get the perfect overhead invention and using 300 caching. Service descriptors have a length of 20 – 25 characters.

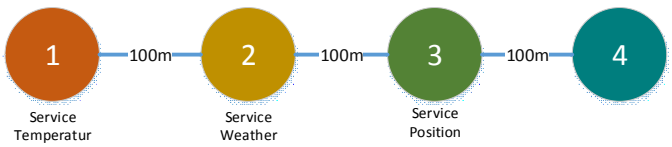


Fig. 8. The bandwidth measurement scenario with node 4 looks for services on nodes 1, 2, 3 and 4

B. Delay Measurement Scenario

The number of hops between nodes that perform *service requests* and *service providers* is the element that holds the most influence on the discovery of the delay of service. To quantify and obtain a time delay, the preferred topology method is a static network node. Nodes connected in node chains 2 through

16 producing 1 – 15 hops as in Fig. 9 [14]. The method used is a single service on a WBN network located at node 0 and requested by the node at the last position of the chain at 20-second intervals. A delay made among an achievement SR and an acceptance mark measured for 100 requests. In the simulation, the test utilizes local caching with a timeout of a 300s timeout. This simulation compares the two conditions of Service Discovery with WBN-P and without WBN-P as a comparison.



Fig. 9. The scenario of delay measurement with some hops N

V. RESULT

A. Bandwidth Measurement

In service discovery on WBN network, to discovery for three service that is weather service, temperature, and position by one node and node number from 4 to 64 shown in Fig.8. The test result, service discovery using WBN-P protocol and without WBN-P protocol at AODV routing protocol gets average efficiency bandwidth 18,6 % if using WBN. The sampling results in bandwidth from the simulations NS-2 shown in Fig. 10. For comparison bandwidth in discovery service graph using WBN-P and without WBN-P on AODV protocol routing as shown in Figure 11.

```
s 0.000169789 _0_ RTR  --- 0 AODV 48 [0 0 0 0] -------- [0:255 -1:255 32 0] [1 0 [HELLO 0 0 0]]
r 0.001478055 _1_ RTR  --- 0 AODV 48 [0 ffffffff 0 800] -------- [0:255 -1:255 32 0] [1 0 [HELLO 0 0 0]]
s 0.044061336 _1_ RTR  --- 1 AODV 48 [0 0 0 0] -------- [1:255 -1:255 32 0] [1 0 [HELLO 1 0 0]]
r 0.045249602 _0_ RTR  --- 1 AODV 48 [0 ffffffff 1 800] -------- [1:255 -1:255 32 0] [1 0 [HELLO 1 0 0]]
r 0.045249602 _2_ RTR  --- 1 AODV 48 [0 ffffffff 1 800] -------- [1:255 -1:255 32 0] [1 0 [HELLO 1 0 0]]
s 0.156480886 _2_ RTR  --- 2 AODV 48 [0 0 0 0] -------- [2:255 -1:255 32 0] [1 0 [HELLO 2 0 0]]
r 0.157969152 _1_ RTR  --- 2 AODV 48 [0 ffffffff 2 800] -------- [2:255 -1:255 32 0] [1 0 [HELLO 2 0 0]]
```

Fig. 10. Results of simulation data with hello packets for bandwidth measurement

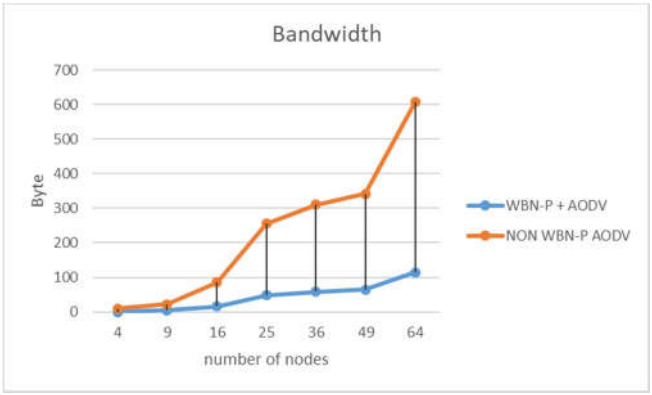


Fig. 11. Bandwidth Measurement Chart

B. Delay Measurement

In service discovery on WBN network, to update data on the local cache on each node as shown in Figure 8. Measurement Service speed discovery with some nodes from 2 to 16. The test results, service discovery using the WBN-P protocol and without the WBN-P protocol on the AODV routing protocol with the number of nodes 4 through 64 received an average efficiency delay of 30.9% when using WBN-P. The sampling results from the simulations shown in Fig. 12. For comparison graph, the delay service discovery uses WBN-P and without WBN-P on AODV protocol routing as shown in Figure 11.

```
s 0.000169789 _0_ RTR  --- 0 AODV 48 [0 0 0 0] -------- [0:255 -1:255 32 0] [1 0 [HELLO 0 0 0]]
r 0.001478496 _4_ RTR  --- 0 AODV 48 [0 ffffffff 0 800] -------- [0:255 -1:255 32 0] [1 0 [HELLO 0 0 0]]
r 0.001478496 _1_ RTR  --- 0 AODV 48 [0 ffffffff 0 800] -------- [0:255 -1:255 32 0] [1 0 [HELLO 0 0 0]]
s 0.044061336 _1_ RTR  --- 1 AODV 48 [0 0 0 0] -------- [1:255 -1:255 32 0] [1 0 [HELLO 1 0 0]]
r 0.045170002 _2_ RTR  --- 1 AODV 48 [0 ffffffff 1 800] -------- [1:255 -1:255 32 0] [1 0 [HELLO 1 0 0]]
r 0.045170043 _3_ RTR  --- 1 AODV 48 [0 ffffffff 1 800] -------- [1:255 -1:255 32 0] [1 0 [HELLO 1 0 0]]
r 0.045170043 _0_ RTR  --- 1 AODV 48 [0 ffffffff 1 800] -------- [1:255 -1:255 32 0] [1 0 [HELLO 1 0 0]]
s 0.082072703 _4_ RTR  --- 2 AODV 48 [0 0 0 0] -------- [4:255 -1:255 32 0] [1 0 [HELLO 4 0 0]]
```

Fig. 12. Results of simulation data with hello packets for delay measurement

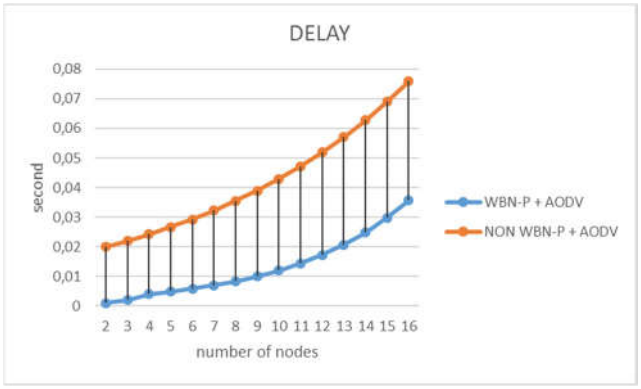


Fig. 13. Delay Measurement Chart

VI. CONCLUSION

In this paper, optimization on Wireless Balloon Network using Breadth Bloom filter that is WBN-P proposed. This method utilizes spare data in the AODV protocol data routing format for service discovery description. Optimization of service discovery description utilizes the Breath Bloom filter (BBf). The results showed that the bandwidth efficiency is 18.6% and the delay effect is 30.9%.

ACKNOWLEDGMENT

This research supported by the Grant of the Ministry of Research, Technology and Higher Education of the Republic of Indonesia.

REFERENCES

[1] Sureshkumar, A., Ellappan, V., & Manivel, K. (2017, March). A comparison analysis of DSDV and AODV routing protocols in mobile AD HOC networks. In *Emerging Devices and Smart Systems (ICEDSS), 2017 Conference on* (pp. 234-237). IEEE.

[2] Xie, Y., Sun, B., Gui, C., & Song, Y. (2017, November). Performance analysis of variable length sliding window network coding in MANETs. In *Systems and Informatics (ICSAI), 2017 4th International Conference on* (pp. 880-884). IEEE..

[3] Liu, X., Li, Z., Yang, P., & Dong, Y. (2017). Information-centric mobile ad hoc networks and content routing: a survey. *Ad Hoc Networks*, *58*, 255-268.

[4] Weng, C. C., Chen, C. W., Ku, C. J., & Hwang, S. F. (2010). A bandwidth-based power-aware routing protocol with low route discovery overhead in mobile ad hoc networks. *The Computer Journal*, *53*(7), 969-990..

[5] Wahyono, I. D. (2016). SERVICE DISCOVERY BERBASIS BREADTH BLOOM FILTER DI MOBILE AD-HOC NETWORK (MANET). *TEKNO*, *23*(1).

[6] Lestari, D., Wahyono, I. D., & Fadlika, I. (2017, November). IoT based Electrical Energy Consumption Monitoring System Prototype: Case study in G4 Building Universitas Negeri Malang. In *Sustainable Information Engineering and Technology (SIET), 2017 International Conference on* (pp. 342-347). IEEE.

[7] Helal, S., Desai, N., Verma, V., & Lee, C. (2003, March). Konark-a service discovery and delivery protocol for ad-hoc networks. In *Wireless Communications and Networking, 2003. WCNC 2003. 2003 IEEE* (Vol. 3, pp. 2107-2113). IEEE.

[8] Zhu, F., Mutka, M. W., & Ni, L. M. (2005). Service discovery in pervasive computing environments. *IEEE Pervasive computing*, *4*(4), 81-90.

[9] Sailhan, F., & Issarny, V. (2005, March). Scalable service discovery for MANET. In *Pervasive Computing and Communications, 2005. PerCom 2005. Third IEEE International Conference on* (pp. 235-244). IEEE.

[10] Stine, J. A. (2006, October). Cross-layer design of manets: The only option. In *Military Communications Conference, 2006. MILCOM 2006. IEEE* (pp. 1-7). IEEE.

[11] Bloom, B. H. (1970). Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, *13*(7), 422-426.

[12] Song, H., Dharmapurikar, S., Turner, J., & Lockwood, J. (2005). Fast hash table lookup using extended bloom filter: an aid to network processing. *ACM SIGCOMM Computer Communication Review*, *35*(4), 181-192.

[13] Rivest, R. (1992). The MD5 Message-Digest Algorithm. RFC 1321 (Informational)

[14] W. Kiess and M. Mauve. A survey on real-world implementations of mobile ad-hoc networks. *Ad Hoc Netw.*, 5(3):324–339, April 2007

[15] University of California. ns2 Network Simulator. (http://www.isi.edu.nsnam/ns/), Accessed 2008

[16] C. Perkins, E. Belding-Royer, and S. Das. Ad hoc On-Demand Distance Vector (AODV) Routing. RFC 3561 (Experimental), July 2003.

Evaluation of Normalization in Fake Fingerprint Detection with Heterogeneous Sensor

Safira Nuraisha
Faculty of Computer Science
Universitas Dian Nuswantoro
Semarang, Indonesia
111201206732@mhs.dinus.ac.id

Guruh Fajar Shidik
Faculty of Computer Science
Universitas Dian Nuswantoro
Semarang, Indonesia
guruh.fajar@research.dinus.ac.id

***Abstract*— Along with the development of a security system model using fingerprint authentication, a new problem arises that is the use of fake fingerprints. Sensors used to capture fingerprint images also influence recognition results. This study aims to detect fake fingerprint images, with a high degree of accuracy, and analyze the effect of normalization on two sensors from different fingerprint images. The dataset used comes from ATVS public data, where the first dataset character uses an optical sensor and a second dataset character using a thermal sensor. Min-max normalization method is proposed to improve the performance of fake fingerprint image detection, with different sensor conditions. To investigate the effect of normalization, we evaluate with the GLCM (Gray Level Co-Occurrence Matrix) image feature, in several different classification methods such as KNN (K-Nearest Neighbor), SVM (Support Vector Machine), and NN (Neural Network). Increased accuracy results obtained in this study with normalization up to 24%.**

Keywords—fake fingerprint detection; GLCM; Min-Max Normalization; Classification.

I. INTRODUCTION

The use of fake fingerprints is a new problem in security system models that use fingerprints. Fingerprint scanners can receive copies of the original fingerprint often called artificial fingerprints. The use of fake fingerprints can threaten the security of a system[1]. The problem of fingerprint detection is getting worse because standard sensors are not able to distinguish images from real fingerprints or fake fingerprints. Second, fake fingerprints are difficult to distinguish from original fingerprints[2].

Another problem, the sensors used to capture the fingerprint image can affect the quality and characteristics of the image[3], it also affects the feature extraction results[4]. After the feature is obtained, the next step is classification process of fake fingerprint and real fingerprint image recognition using two different datasets. First datasets is use flat optical sensors to capture the image, second, use thermal sweeping sensors. These differences can be identified based on the gray level of the fingerprint image[5]. The proposed feature extraction using the Gray Level Co-Occurrence Matrix (GLCM) method. GLCM is one of the most popular texture-based feature extraction methods. Due to its effectiveness in texture analysis, GLCM is very effective in recognition analysis [2][6][7][8][9].

After the feature extraction process, the next process is normalization. In this study, the normalization method that used is Min-Max Normalization. It normalized the feature extracted data. The min-max method is chosen in this study because the data characteristics have a considerable range of differences, so by using this method, the range of data with the highest and lowest values can be changed to produce normal data.

This study aims to analyze the effect of normalization on use of two fingerprint image sensors for detecting fake fingerprint images with a high degree of accuracy

The discussion in this study organized as follows: Section II describe the related work of this study. Section III show the method and evaluation process that use in this study. Section IV describe about result and analysis of the proposed method. The last section is the conclusion.

II. RELATED WORK

Previous research by Ankita Chaudhari and P.J. Deore [10] aims to compare two images of original fingerprints or fingerprints. the data used comes from one user. The result of this study is match between the fake fingerprint and real fingerprint. Using various features related to histogram, minutiae extraction and GLCM, Ankita Chaudhari and P.J. Deore can find a model proposal that will recognize the spoof attack. The percentage of FAR yields 92.44% while the FRR is 92.30%.

The study of three artificial fingerprint databases was evaluated using SVM classification techniques with different sample and feature vectors [11]. The results of the study showed that latex-based fingerprints are more difficult to evaluate their authenticity. Databases derived from CASIA and Livdet13 can be easily evaluated using the standard algorithm of artificial fingerprint classification.

Previous research on coconut wood quality classification method using Gray Level Co-occurrence matrix extraction method[9]. This study discusses the classification of quality of coconut wood. The method of feature extraction used to get character from coconut wood is GLCM. The classification method used are AutoMLP and SVM. Based on the experimental results, the best method used for the classification is AutoMLP. The accuracy up to 78.82%.

III. METHOD

A. Gray Level Co-Occurance Matrix(GLCM)

GLCM is one of the most popular extraction texture methods in research on image[6]. Information related to the relative position of neighboring pixels obtained with GLCM. This method is very effective in texture analysis and image recognition[7]. GLCM consists of a two-dimensional histogram, which in each element (m, n) is the frequency of occurrence m along with the incidence of n. The co-occurrence matrix is determined by the relative frequency P (m, n, d, θ) in which two pixels are separated by angle θ, the first with the gray m and the other with the gray level n. the co-occurrence matrix show in equation 1, is determine

$$\sum_{x=1}^{N}\sum_{y=1}^{N}\begin{cases}1, if\ I(x,y)=m\ and\ I(x+\Delta x, y+\Delta y)=n\\ 0, otherwise\end{cases} \quad (1)$$

Where Δx and Δy are offsets and specify the distance of x and y between the pixel-of-interest and its neighbors. Some studies with GLCM calculate different orientations of the degrees which can extract reliable features from images and illustrate the spatial relationships of neighboring pixels. Haralick[12] proposed 14 texture features, Haralick claims that each of the 14 features provides clues about some of the textual characteristics of an image.

B. Min-Max Normalization

There are several methods that can be used to normalize the data set, which includes normalized min-max, z-score normalization, decimal scales, standard time etc. Min-Max Normalization method is a method that applied quite a lot in research related to fingerprint so, one of them in stage improvement of image quality[1]. In this study, the min-max normalization method is used to normalize the data before going through the classification process. Normalization using min-max method, leveling the distance of attribute value a based on its minimum and maximum value. This method changes the value of a to a' by calculating the range of low to high[13], that show in equation 2:

$$a' = (high - low)\ x\ \frac{a-minA}{maxA-minA} + low \quad (2)$$

a '= normalization result data
low = minimum value given
high = maximum value given
minA = the minimum value of data per column
maxA = maximum value of data per column

C. Classification

One of the best machine learning methods that can solve the classification problem is Support Vector Machine [14]. SVM is often used in image processing and pattern recognition applications because of its computational efficiency, rapid convergence and exceptional generalization capabilities and so on. The main purpose of SVM is to find the optimal separator surface, which is a hyperplane that is equidistant from two classes, to separate it[15].

KNN is one method of classification with pattern recognition based on statistics[6]. The stages in the process of this classification are determine the number of k, which is the number of closest classes. Then calculate the distance between the new image of the train data. Specifies the nearest neighbor by sorting the distance according to the minimum distance k. Check the label in the closest class. Finally, classify the test image to the closest class.

The first stage in Neural Network method is initialization of initial value and bias value. Then calculate the input, output, and error values for each node. To generate output value, use sigmoid activation function.

D. Evaluation Process

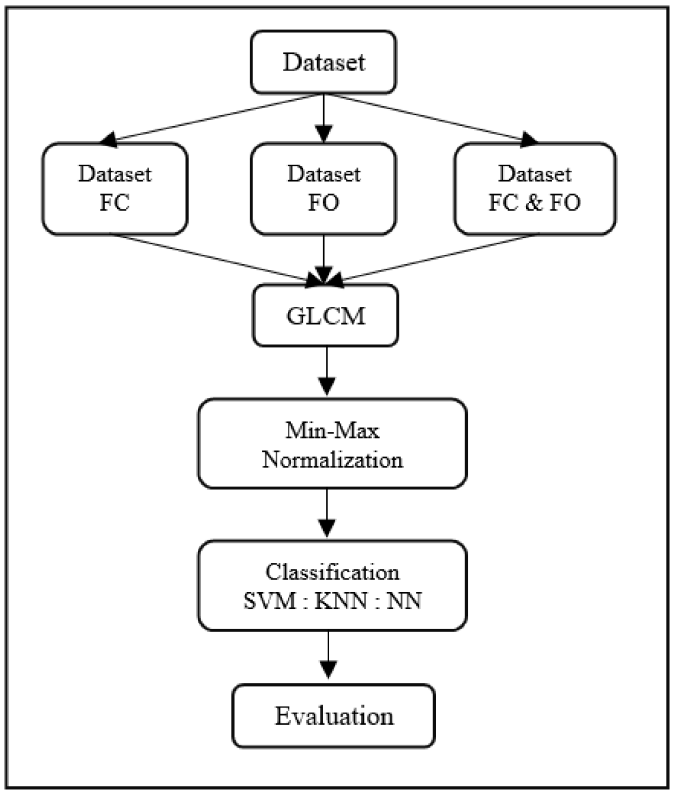


Fig 1. Evaluation Process

There are five main stages in the process of fingerprint image recognition and analysis of the effect of normalization on fingerprint image dataset from different sensors.

Dataset

The dataset used is a public dataset of the original fingerprint image or artificial fingerprints obtained from Biometrics and Data Pattern Analytics - ATVS Biometric Recognition Group[16]. The dataset is an image with a grayscale pixel color consisting of a fingerprint sample of the forefinger and middle finger were taken in one acquisition session with:

- Flat optical sensor (512 dpi) which is then called "*dataset fc*"
- Thermal sweeping sensor (500 dpi) which is then called "*dataset fo*"

This dataset consists of 68 fingers x 4 samples x 2 sensors = 544 samples of the original fingerprint image and 544 fingerprint images, so, the total data is 1088 image.

Preprocessing

The preprocessing stage divides the dataset based on the sensor used to retrieve the fingerprint image. There are two sensors used are biometric flat optical sensors which are then called "fc" and thermal sweeping sensors which are then called "fo". The data is divided into three categories; fc, fo, and the combination between fc and fo.

Feature Extraction using GLCM

This research using GLCM as a feature extraction method [6]. Stage performed for feature extraction with GLCM is find matrix in four orientations of 0, 45, 90 and 135 degrees each for the images with distances of one pixel. In this research, follow the best existing work GLCM, using contrast feature[1], the equation of contrast feature shown bellow:

$$\text{Contrast} = \sum_{n=0}^{Ng-1} n^2 \{ \sum_{i=1}^{Ng} \sum_{j=1}^{Ng} p(i,j) \} \quad (3)$$

Data Normalization using Min-Max Normalization

Steps after feature extraction are then normalized the data. Characteristics of fc dataset images and fo datasets are very different, so feature extraction results in data have much different range. Normalization is useful to equate the range of feature extracted data from datasets with different sensors.

Classification

After the dataset is normalized, the next step is classification. The classification in this study uses the Support Vector Machine method with kernel dot type and C = 0 value, then compared with two methods of classification of KNN and Neural Network.

Evaluation

Calculation of performance in this study using accuracy. Experimental testing and validation is the result of the introduction of test data against the system.

Accuracy is calculated to find the level of accuracy, which is formulated as follows:

$$\text{Accuracy} = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \quad (4)$$

To calculate the increase in accuracy due to the effect of normalization, the following calculations are:

$$\text{Increase In Accuracy} = \frac{A - A'}{A} * 100\% \quad (5)$$

A = accuracy value without normalization
A' = accuracy value using normalization

IV. RESULT AND ANALYSIS

After the experiment, the experimental results are described as follows.

Preprocessing Result

After the experiment, the results are described as follows. The original datasets used in this study consisted of fingerprint images with two different sensors.

TABLE I. SAMPLE DATASET[16]

No	File Name	Sensor	Sample	Label
1	u01_f_fc_li_03	Optical		Fake
2	u01_o_fc_li_03	Optical		Real
3	u01_f_fo_li_03	Thermal		Fake
4	u01_o_fo_li_03	Thermal		Real

Then the datasets are grouped according to the real image fingerprint and fake fingerprints.

Feature Extraction using GLCM Result

The feature extraction data using GLCM consists of four attributes consisting of a contrast value of 0 degrees orientation, a contrast value of 45 degrees orientation, a contrast value of 90 degrees orientation, and a 145 degrees orientation contrast, as well as a label attribute. Examples of data as follows.

TABLE II. EXAMPLES DATA AFTER FEATURE EXTACTION

Data	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Class
Image 1	3907	5	2	4	0
Image 2	9215	8	4	20	0
Image 3	23	24	12	2	1
Image 4	741	17	20	12	1

Data Normalization using Min-Max Normalization Result

After the data obtained from feature extraction, normalize process is done to generate data with a range of 0 to 1 data could be show in table 3

TABLE III EXAMPLES DATA AFTER NORMALIZATION

Data	Attribute 1	Attribute 2	Attribute 3	Attribute 4	Class
Image 1	0.325	0.088	0.000	0.083	0
Image 2	0.769	0.176	0.083	0.417	0
Image 3	0.001	0.647	0.417	0.000	1
Image 4	0.061	0.441	0.750	0.750	1

Classification and Evaluation Result

After the normalization, we evaluate the image features using three classifiers, there are KNN, SVM, and Neural Network. We use parameter kernel dot in SVM. In KNN we use neighbor k = 1 and euclidian distance. After that, the NN we use two hidden layers with feed forward neural network. After we class the data, we calculate using confussion matrix.

The test is divided into 3 stages: experiment with dataset fc, experiment with fo dataset, and experiment with dataset fc and fo.

TABLE IV. RESULT

Classifier	Datasets	Without Normalization	With Normalization
KNN	fo	86.41%	88.62%
	fc	80.34%	80.89%
	fo and fc	83.64%	84.93%
SVM	fo	93.21%	93.21%
	fc	84.93%	84.93%
	fo and fc	61.77%	76.85%
NN	fo	98.54%	98.54%
	fc	88.05%	88.05%
	fo and fc	81.44%	81.72%

Comparison of accuracy of data before normalized and after normalization are perform in Table IV. The normalization gives affects to classification results. In a single sensor dataset, the classification performance before and after normalization have not significant result. In two combined sensors dataset, the normalization able to improve the accuracy performance significantly in SVM, KNN, and NN up to 24%, 1.5% and 0.34% sequentially

CONCLUSION

This study has shown that fingerprint forgery can be detected through well-proposed methods. The sensors used in fingerprint image retrieval greatly affect the feature extraction results. The application of normalization phase on feature extraction result gives improve accuracy of classification with Support Vector Machine, KNN, and NN method up to 24%, 1.5% and 0.34%. The future work, it necessary features extraction that able to contrast the fake fingerprint feature thus improve the classification performance.

REFERENCE

[1] D. V. L. P. V. Sasikala, "A Comparative Study on the Swarm Intelligence Based Feature Selection Approaches for Fake And Real Fingerprint Classification," *Int. Conf. Soft-Computing Netw. Secur.*, 2015.

[2] A. Rattani, Z. Akhtar, and G. L. Foresti, "A Preliminary Study on Identifying Fabrication Material from Fake Fingerprint Images," *IEEE Symp. Ser. Comput. Intell. A*, pp. 362–366, 2015.

[3] A. Rattani, W. J. Scheirer, and A. Ross, "Open Set Fingerprint Spoof Detection Across Novel Fabrication Materials," *IEEE Trans. Inf. FORENSICS Secur.*, vol. 10, no. 11, pp. 2447–2460, 2015.

[4] G. F. Shidik, F. N. Adnan, C. Supriyanto, R. A. Pramunendar, and P. N. Andono, "Multi Color Feature , Background Subtraction and Time Frame Selection for Fire Detection," *Int. Conf. Robot. Biomimetics, Intell. Comput. Syst.*, no. November, pp. 115–120, 2013.

[5] K. Abhishek and A. Yogi, "A Minutiae Count Based Method for Fake Fingerprint Detection," *Procedia Comput. Sci.*, vol. 58, pp. 447–452, 2015.

[6] A. Ali, X. Jing, and N. Saleem, "GLCM-Based Fingerprint Recognition Algorithm," *IEEE IC-BNMT*, no. 207–211, 2011.

[7] A. Eleyan and H. Demirel, "Co-Occurrence based Statistical Approach for Face Recognition," *ISCIS*, vol. 0, pp. 611–615, 2009.

[8] T. Cevik, A. M. A. Alshaykha, and N. Cevik, "Performance Analysis ofGLCM-based Classification on Wavelet Transform-Compressed Fingerprint Images," *Digit. Inf. Commun. Technol. its Appl.*, pp. 131–135, 2016.

[9] R. A. Pramunendar, C. Supriyanto, D. H. Novianto, I. N. Yuwono, G. F. Shidik, and P. N. Andoro, "Classification Method of Coconut Wood Quality," *Int. Conf. Robot. Biomimetics, Intell. Comput. Syst.*, no. November, pp. 254–257, 2013.

[10] A. Chaudhari and P. J. Deore, "Prevention of spoof attacks in fingerprinting using histogram features," *NIRMA Univ. Int. Conf. Eng. NUiCONE*, pp. 1–4, 2012.

[11] Q. Huang, S. Chang, C. Liu, B. Niu, M. Tang, and Z. Zhou, "An evaluation of fake fingerprint databases utilizing SVM classification," *Pattern Recognit. Lett.*, vol. 60–61, pp. 1–7, 2015.

[12] R. M. Haralick and K. Shanmugam, "Textural Features for Image Classification," *IEEE Trans. Syst. MAN, Cybern.*, pp. 610–621, 1973.

[13] E. Ogasawara, L. C. Martinez, D. De Oliveira, G. Zimbrão, G. L. Pappa, and M. Mattoso, "Adaptive Normalization : A Novel Data Normalization Approach for Non-Stationary Time Series," *Neural Networks (IJCNN)*, 2010.

[14] V. N. Vapnik, "Statistical Learning Theory, Wiley, New York," 1998.

[15] L. Boutella, A. Serir, and T. H. Boumediene, "Fingerprint Identification by Wave Atoms Transform and SVM," *Int. Conf. Adv. Syst. Electr. Technol.*, pp. 301–306, 2017.

[16] J. Galbally, S. Marcel, and J. Fierrez, "Image quality assessment for fake biometric detection: Application to Iris, fingerprint, and face recognition," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 710–724, 2014.

Pulse Width Modulation (PWM) and Pulse Amplitude Modulation (PAM) Technique for Medium-Speed BLDCM in Electric Vehicle Application

Muhammad Rizani Rusli[1], Haris Rahmatullah [2], Melinda Badriatul Fauziah [3], Arman Jaya [4], Mochammad Machmud Rifadil[5], and Era Purwanto[6]
[1,4,6]Postgraduate of Electrical Engineering
[2,3,4,5,6]Industrial Electrical Engineering
Electrical Engineering Department, Politeknik Elektronika Negeri Surabaya (PENS), Surabaya, Indonesia
[1]ruslirizani@gmail.com, [2]harisrahmatullah19@gmail.com, [3]melfauziah@gmail.com, [4]arman@pens.ac.id, [5]mmrifadil@pens.ac.id, [6]era@pens.ac.id

***Abstract*—High torque and high efficiency are two factors used as a consideration of why many electric vehicles (EV) use brushless DC motor (BLDCM) as a propulsion system. Likewise, Amtenar car —a PENS's electric car— use this kind of motor as its propulsion. BLDCM can be driven by two techniques. They are Pulse Amplitude Modulation (PAM) and Pulse Width Modulation (PWM) technique. Both techniques are applied to the Amtenar car using the STM32F4 microcontroller to observe the dynamic speed response and efficiency which is operated under no-load conditions and on a load of Amtenar car. PWM scheme uses H-PWM-L-ON while PAM technique employes DC-DC buck converter. When PAM technique is applied to a no-load condition, PAM technique produces higher speed and efficiency and also produces a better dynamic speed performance as compared to PWM technique. While, when both techniques are applied to Amtenar car in the same line-to-line RMS voltage reference, PAM technique generates higher speed and has better dynamic speed performance than PWM technique. According to the comparative study of two techniques that have been observed in this paper, an efficient and high-performance technique for the BLDCM drive as an EV propulsion can be determined.**

Keywords— Brushless DC Motor (BLDCM), Pulse Width Modulation (PWM), Pulse Amplitude Modulation (PAM), Electric Vehicle.

I. INTRODUCTION

Nowadays, EV technology has increased in the research and development. Many types of EV use BLDCM as a propulsion system because it has advantages of high efficiency, high driving torque at low speed, high power density, low maintenance, and high dynamic speed [1] [2] [3] [4].

BLDCM requires a DC power supply that is converted to AC voltage through a 3-phase inverter with a commutation based on the rotor position. PAM and PWM are two commonly used in BLDCM drive techniques [5].

PWM technique requires constant DC input voltage which is chopped with constant frequency and adjustable duty cycle according to speed and load. PWM technique is often used in BLDCM drives because it is simple without requiring an additional DC-DC converter. PWM techniques have disadvantages such as increasing the thermal loss because of the magnetic hysteresis [6], having a large switching loss because of applying PWM signal to MOSFET of 3-phase inverter [7], and increasing iron and copper loss at high speed [5]. There are 6 PWM schemes on the BLDCM drives —H-ON-L-PWM, H-PWM-L-ON, PWM-ON, ON-PWM, PWM-ON-PWM, and H-PWM-L-PWM—. PWM-ON-PWM, PWM-ON, and H-PWM-L-ON have the lowest torque ripple on the commutation as compared to other PWM schemes [4].

On the PAM technique, the output voltage of the 3-phase inverter is controlled by adjusting the DC-link voltage, requiring a DC-DC converter on the input side of the 3-phase inverter. Some types of DC-DC converter used in many papers are buck converter [8] [9], buck-boost converter [10] [11], bridgeless PFC buck-boost converter [12], cascade buck-boost converter [13], cuk converter [14], etc. The type of the DC-DC converter which is often used in some papers and applications is buck converter. Buck converter has a simple structure and also has an inductor and capacitor working to produce a smooth DC output voltage. In addition, it can reduce noise caused by hard switching of PWM [15]. The commutation of inverter switching is 120° commutation or six-step commutation according to the hall effect sensor used to detect the rotor position [5].

There are also several studies related to the comparison between PAM and PWM in some papers. In [5], they discuss the comparison of PAM and PWM techniques on high speed and high voltage BLDCM with detailed theoretical analysis and analyze the THD current, power factor, and efficiency. Wu, Genfu, and Ying discuss the comparison between PAM and PWM on BLDCM by using adaptive lines and analyze the power factor, THD current, and efficiency in [16]. In [17], they discuss the efficiency between PAM and PWM of sensorless BLDCM in refrigerator applications. In [5], [16] and [17], they use main power coming from AC source so that they use PFC. Kim and Youn compare the performance of

PAM and PWM of sensorless high-speed BLDCM by observing commutation delay at high-speed operation in [7]. In [18], they discuss the comparative performance of electromagnetic and power losses on SPM BLDCM for turbocharger applications with PAM and PWM techniques.

This work examines a comparative study of two techniques (PAM and PWM) and observes under the no-load condition and with a load of Amtenar car using the STM32F4 microcontroller. PWM scheme uses H-PWM-L-ON while PAM technique employes DC-DC buck converter. This paper analyses the dynamic speed response and efficiency of the BLDCM.

II. BLDCM

BLDCM is one kind of the 3-phase AC electric synchronous motors having a trapezoidal back electromotive force (BEMF). Conventional DC motor uses a brush for commutation process while BLDCM uses the electrical contact for the commutation process because the winding of the armature and the permanent magnet is placed on the stator and rotor side separately. So, the rotation and direction of the flux are controlled by hall effect sensor as a rotor position detector used to determine the commutation of 3-phase inverter switching [19] [20].

BLDCMs are modeled by assuming that wye connection is applied on stator's winding, the value of each phase's resistance is equal, and the iron core losses are ignored [19]. The equivalent circuit of BLDCM is shown in Fig. 1.

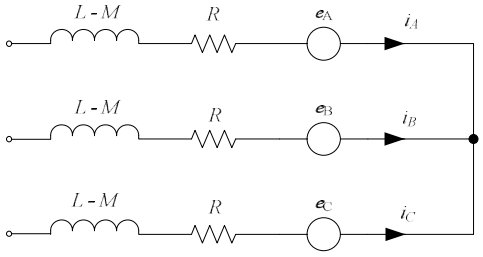


Fig. 1. The equivalent circuit of BLDCM

The equation of BLDCM each phase is written in (1).

$$u_A = Ri_A + (L-M) + \frac{d}{dt} e_A \quad (1)$$

Then, the matrix form of the phase voltage equation on the BLDCM can be written in (2).

$$\begin{bmatrix} u_A \\ u_B \\ u_C \end{bmatrix} = \begin{bmatrix} R & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & R \end{bmatrix} \begin{bmatrix} i_A \\ i_B \\ i_C \end{bmatrix} + \begin{bmatrix} L-M & 0 & 0 \\ 0 & L-M & 0 \\ 0 & 0 & L-M \end{bmatrix} \frac{d}{dt} \begin{bmatrix} i_A \\ i_B \\ i_C \end{bmatrix} + \begin{bmatrix} e_A \\ e_B \\ e_C \end{bmatrix} \quad (2)$$

Where, u_A, u_B, u_C, are the voltage at phases *A, B,* and *C*. *R* is the stator resistance assumed to be equal, *L* is self-inductance and *M* mutual inductance. i_A, i_B, i_C, are the currents in phase *A*, *B*, and *C*. e_A, e_B, and e_C, are BEMF voltages in phases *A, B,* and *C*.

The magnitude of the torque on the BLDCM can be written in (3).

$$T_e = \frac{e_A i_A + e_B i_B + e_C i_C}{\omega_m} \quad (3)$$

Where,

T_e = Electromagnetic torque (N.m)
ω_m = Angular velocities (rpm)

The speed of the BLDCM is given in (4).

$$T_e - T_L = J\frac{d\omega_m}{dt} + B_v\omega_m \quad (4)$$

Where,

T_L = Load torque (N.m)
J = Moment of rotor inertia ($Kg.m^2$)
B_V = Coefficient of friction

The relationship between the rotor speed and the rotor position is expressed as (5).

$$\frac{d\theta_r}{dt} = \frac{P}{2}\omega_m \quad (5)$$

Where,

θ_r = the position of the rotor based on d-axis (rad)
P = the number of poles.

The hall-effect sensor on BLDCM is used to detect the position of the rotor that will be used to determine the commutation of the 3-phase inverter. The magnetic field will be generated and the rotor will rotate when the current flows through the coil.

III. BLDCM DRIVES

A. PWM Technique

One of the popular BLDCM drives is PWM technique. When the commutation switch of the inverter is ON, it is divided into several pulses which have a higher frequency than the six-step switching frequency of the inverter. The width of the pulse can be modulated in which it makes the output voltage of 3-phase inverter with constant DC input voltage can be adjusted [17].

Block diagram of the BLDCM drive with PWM technique is shown in Fig. 2.

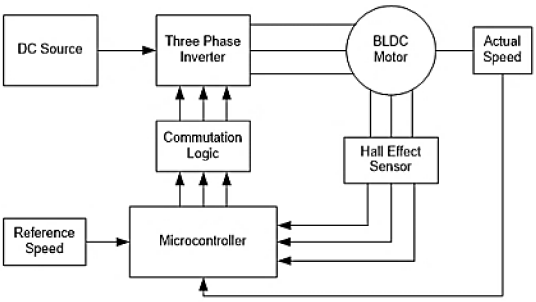


Fig. 2. Block diagram of the BLDCM drive with PWM technique

The six-step inverter switching commutation scheme with PWM technique is depicted in Fig. 3.

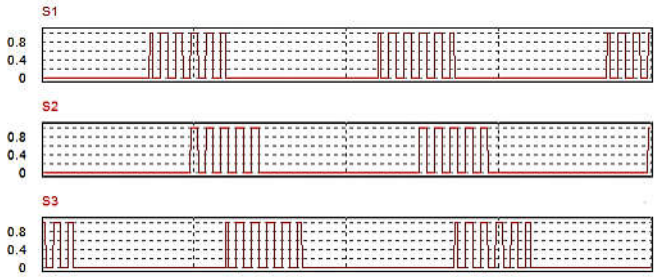


Fig. 3. Switching sequence of the 3-phase inverter with PWM technique

B. *PAM Technique*

The BLDCM drive can also be done by PAM technique. The commutation of the inverter's switching is 120 or six-step commutation according to the hall-effect sensor which detects the position of the rotor. By changing the DC-link voltage, the output voltage of the inverter is adjusted. In this paper, the DC-link voltage of the inverter is adjusted by applying a DC-DC buck converter with Continuous Conduction Mode (CCM) design. The range of DC-link voltage starts from 0V to 48V. Table I shows the kinds of parameters are used in DC-DC buck converter.

TABLE I. THE SPECIFICATION OF DC-DC BUCK CONVERTER

Vin min	46.04
Vin max	52.2 V
Vo	48 V
Io	20 A
Duty Max	91%
L	48.39 μH
C	150 μF

Process flowchart of the BLDCM drive with PAM technique is shown in Fig. 4.

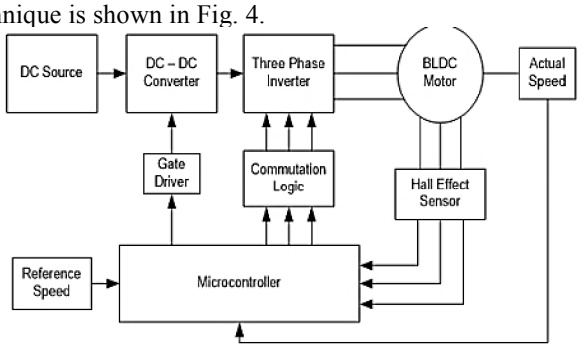


Fig. 4. Process flowchart of the BLDCM drive with PAM technique

The six-step inverter switching commutation scheme with PAM technique is depicted in Fig. 5.

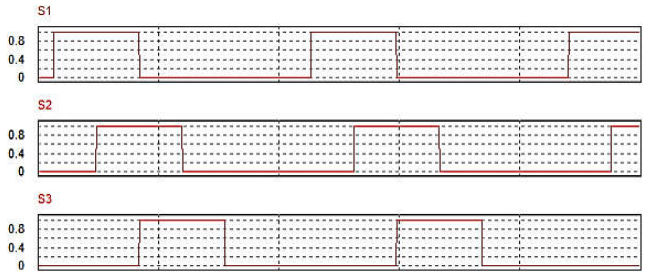


Fig. 5. Switching sequence of the 3-phase inverter with PAM technique

IV. Experimental Result and Discussion

This paper is completed with a hardware implementation. The motor used is BLDCM 1 kW 48V. The hardware of this system are the microcontroller board STM32F4VGTx, IR2110 3-phase inverter driver, FOD3182 MOSFET gate driver for a buck converter, buck converter to vary the DC-link voltage, and 3-phase inverter with PWM scheme of H-PWM-L-ON to drive the BLDCM. The experimental setup is shown in Fig. 6.

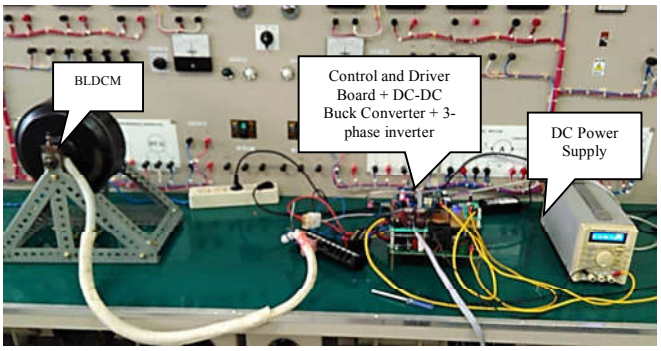
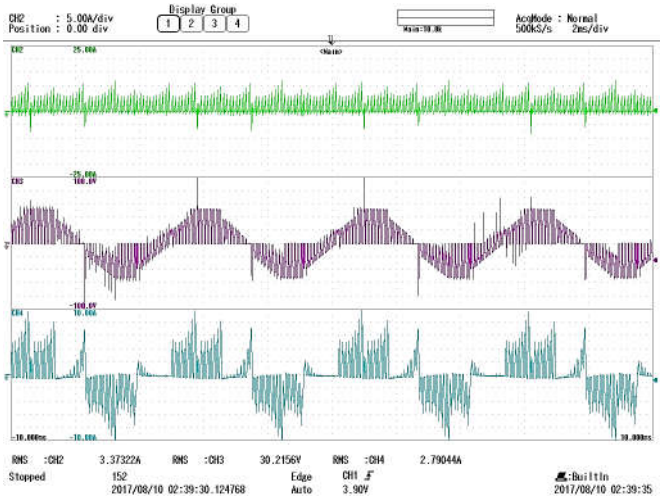
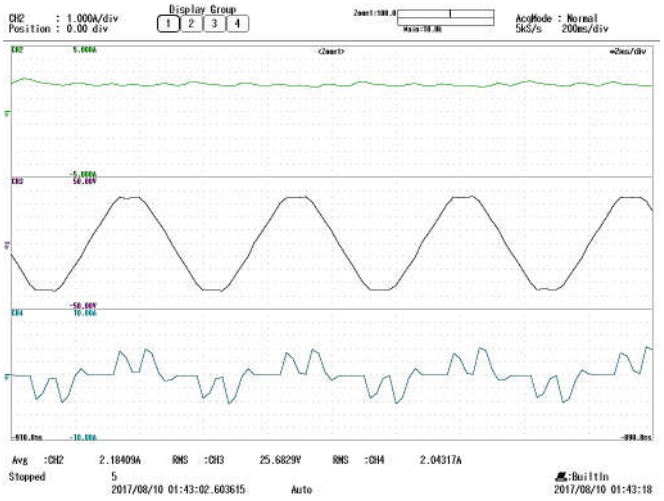


Fig. 6. Experimental setup in no-load operation

In this paper, the parameters observed are input and output of 3-phase inverter voltage and current, input and output power from BLDCM, and also dynamic speed response of BLDCM.



(a)



(b)

Fig. 7. The waveforms of measurement on each technique including DC input current, BEMF line-to-line A-B voltage, and phase A current respectively. (a) PWM technique. (b) PAM technique.

Fig. 7 depicts the result for measuring the DC input current, BEMF line-to-line AB voltage and phase current waveforms of PWM and PAM technique at 500 rpm. As shown in Fig. 7, the waveforms of DC input current in PAM technique have a smaller ripple than PWM technique has. Table II and III show the result of this measurements.

TABLE II. EXPERIMENTAL RESULT OF PWM TECHNIQUE

Speed (rpm)	Vdc (V)	Idc (A)	Vll A-B rms (V)	Iac rms (A)	P in (watt)	P out (watt)	Efficiency (%)
100	53	1.16	12.69	1.53	61.62	33.72	54.73
200	53	1.87	16.92	1.93	99.12	56.66	57.16
300	53	2.42	20.9	2.31	128.21	83.52	65.14
400	53	3.01	26.03	2.75	159.62	124.13	77.76
500	53	3.37	30.22	2.79	178.78	146.04	81.69

TABLE III. EXPERIMENTAL RESULT OF PAM TECHNIQUE

Speed (rpm)	Vdc (V)	Idc (A)	Vll A-B rms (V)	Iac rms (A)	P in (watt)	P out (watt)	Efficiency (%)
100	53	0.61	5.45	1.18	32.33	11.12	34.41
200	53	0.96	10.28	1.93	50.68	36.23	71.48
300	53	1.37	15.72	2.43	72.58	66.10	91.08
400	53	1.75	20.72	2.10	92.87	75.27	81.05
500	53	2.08	26.15	2.08	110.19	94.24	85.53

Regarding Table II and III, it is clear that overall results of DC input current in PAM technique are smaller than PWM technique. The result of the output voltage and current in PAM technique is also smaller than PWM technique. Thus, the input and output power of each technique can be calculated so it can be used to calculate the efficiency. The efficiency of PAM technique is greater than that of PWM technique at speed of 200 rpm up to 500 rpm and it can be shown on the curve in Fig. 8.

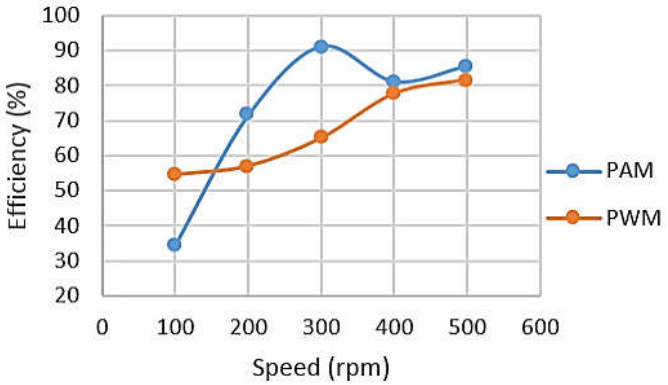


Fig. 8. Efficiency comparison of PWM and PAM techniques with no-load

To observe the dynamic speed response of PAM and PWM technique, the system was tested at 200 rpm and 400 rpm which are shown in Fig. 9 and Fig. 10.

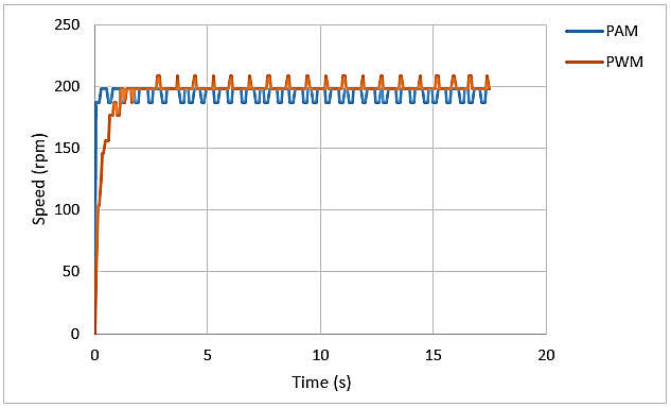


Fig. 9. Speed response of BLDCM with no-load at 200 rpm

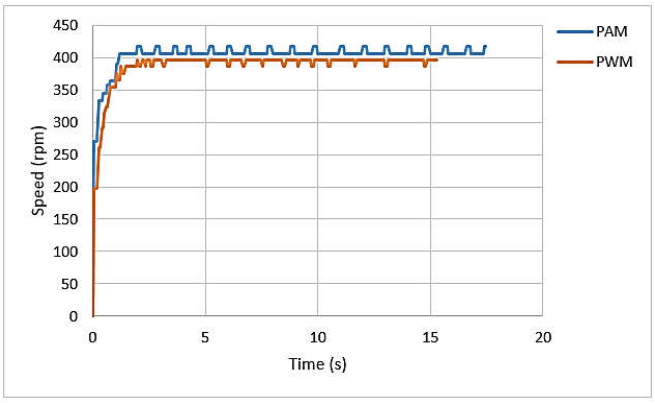


Fig. 10. Speed response of BLDCM with no-load at 400 rpm

According to the experimental result in Fig. 9, it can be seen that when using PWM technique at 200 rpm, the result shows 0.96 seconds as the rise time and 1.2 seconds as the settling time. While using PAM technique at 200 rpm, the rise time is 0.208 seconds and the settling time is 0.26 seconds. Then, the experiment is also tested at 400 rpm as shown in Fig. 10. It can be seen that the result 1.008 seconds as a rising time and 1.26 seconds as a settling time when using PWM technique. While using the PAM technique, it has a rise time of 0.96 seconds and settling time of 1.2 seconds.

The overall result of the BLDCM dynamic speed performance with the no-load condition is shown in Table IV. Where Tr, Ts, and Mp stand for rising time (s), settling time (s) and overshoot (%) respectively. It can be concluded that the speed response of PAM technique has a smaller rising time and settling time value than PWM technique. Furthermore, the result shows that there is zero overshoot along with the experiments.

TABLE IV. DYNAMIC SPEED PERFORMANCE OF PAM AND PWM TECHNIQUE UNDER NO-LOAD CONDITION

No.	Speed (rpm)	PWM Control Technique			PAM Control Technique		
		Tr (s)	Ts (s)	Mp (%)	Tr (s)	Ts (s)	Mp (%)
1.	150	0.84	1.06	0	0.208	0.26	0
2.	200	0.96	1.2	0	0.208	0.26	0
3.	300	1.48	1.86	0	1.168	1.46	0
4.	400	1.008	1.26	0	0.96	1.2	0
5.	500	0.808	1.01	0	0.264	0.33	0

The next experimental test uses a load of Amtenar car —an EV with the 208 kg of weight, single-seat, powered by 1kW BLDCM, and made by PENS' students— which can be seen in Fig. 11. The test is performed by referring the same 3-phase inverter output voltage to compare the dynamic speed performance of the PAM and PWM technique.



Fig. 11. Amtenar car

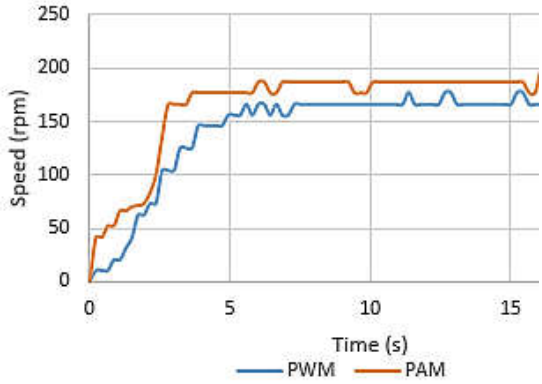


Fig. 12. Speed response of BLDCM with load (at Vline-to-line A-B 10V)

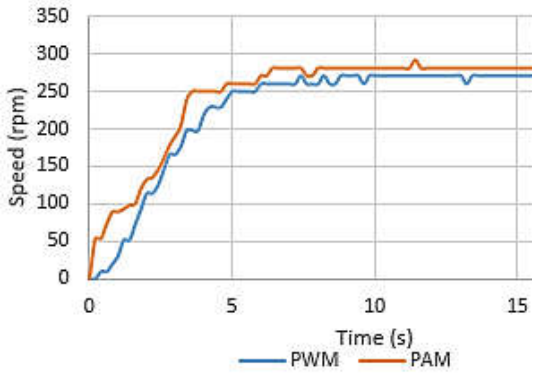


Fig. 13. Speed response of BLDCM with load (at Vline-to-line A-B 15V)

Fig. 12 and Fig. 13 depict the speed response of BLDCM with load condition when the reference is Vline-to-line A-B 10V and 15V respectively. For the more results, it can be shown in Table V.

TABLE V. DYNAMIC SPEED PERFORMANCE OF PAM AND PWM TECHNIQUE WITH LOAD CONDITION.

Vline-to-line A-B (V)	PWM Technique				PAM Technique			
	Tr (s)	Ts (s)	Mp (%)	Steady state speed (rpm)	Tr (s)	Ts (s)	Mp (%)	Steady state speed (rpm)
5	3.25	4	0	83	1.5	3.5	0	93
10	3.64	4.9	0	160	2.5	3.6	0	177
15	4	7.4	0	271	3.4	6.4	0	281

The experimental result with a load of Amtenar car is shown in Table V. It can be concluded that by using PAM technique, the rising time and settling time would be shorter than that of PWM technique. In addition, with the same line-to-line voltage supply to the BLDCM, the steady-state rpm value of PAM technique is higher than that of PWM technique.

V. CONCLUSION

BLDCM drive based on the experimental result with PWM and PAM techniques have been presented. When applied to a no-load condition, PAM technique produces higher speed and efficiency and also produces a better dynamic performance as compared to PWM technique. Moreover, when applied to Amtenar car in the same line-to-line RMS voltage reference, PAM technique produces higher speed and better dynamic speed performance than PWM technique. With the comparative study of two techniques that observed in this paper, it can be determined that PAM technique has an efficient technique and a high performance on BLDCM drive when applied in EV to obtain a high acceleration. For the further research, it had a better focus on the improvement of the types of the DC-DC converter which enable to handle a wide output voltage range with high efficiency and PWM schemes on the 3-phase inverter in order to determine higher efficient and higher performance drive.

ACKNOWLEDGMENT

This research would not finish without the support of Electric Drive Laboratory and Electric Vehicle Community (EVC) PENS due to the academic guidance and providing the facilities to finish this research. The authors also would like to say happy 10th anniversary for EVC-PENS, do the best for contributing the research and development of EV in Indonesia.

REFERENCES

[1] A. Jaya, E. Purwanto, M. B. Fauziah, F. D. Murdianto, G. Prabowo and M. R. Rusli, "Design of PID-fuzzy for speed control of brushless DC motor in dynamic electric vehicle to improve steady-state performance," in *2017 International Electronics Symposium on Engineering Technology and Applications (IES-ETA)*, Surabaya, 2017.

[2] A. Jaya, E. Wahjono, M. R. Rusli, E. Purwanto, F. D. Murdianto and M. B. Fauziah, "Design and simulation of sensorless BLDC motor drive using flux linkage increment based on the line-to-line BEMF for electric vehicles," in *2017 International Electronics Symposium on Engineering Technology and Applications (IES-ETA)*, Surabaya, 2017.

[3] C.-l. Xia, Permanent Magnet Brushless DC Motor Drives and Controls, Wiley, 2012.

[4] S. Mozaffari Niapour, G. Shokri Garjan, M. Shafiei, M. Feyzi, S. Danyali and M. Bahrami Kouhshahi, "Review of Permanent-Magnet Brushless DC Motor Basic Drives Based on Analysis and Simulation Study," *International Review of Electrical Engineering (IREE),* vol. 9, no. 5, pp. 930-957, 2014.

[5] C. Cui, G. Liu, and K. Wang, "A Novel Drive Method for High-Speed Brushless DC Motor Operating in a Wide Range," *IEEE Transactions on Power Electronics,* vol. 30, no. 9, pp. 4998-5008, 2015.

[6] J. Musil, "3-Phase BLDC Drive Using Variable DC Link Six-Step Inverter Designer Reference Manual," Freescale Semiconductor, 2006.

[7] K.-H. Kim and M.-J. Youn, "Performance comparison of PWM inverter and variable DC link inverter schemes for high-speed sensorless control of BLDC motor," *Electronics Letters,* vol. 38, no. 21, pp. 1294-1295, 2002.

[8] Z. Xiaofeng and L. Zhengyu, "A New BLDC Motor Drives Method Based on BUCK Converter for Torque Ripple Reduction," in *2006 CES/IEEE 5th International Power Electronics and Motion Control Conference*, Shanghai, 2006.

[9] J. Feng, K. Liu and Q. Wang, "Scheme based on buck-converter with three-phase H-bridge combinations for high-speed BLDC motors in aerospace applications," *IET Electric Power Applications,* vol. 12, no. 3, pp. 405-414, 2018.

[10] M. Banagar and A. Usha, "Design and simulation of bridgeless PFC buck-boost converter fed BLDC motor drive," in *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Vellore, 2017.

[11] R. Kumar and B. Singh, "Buck-boost converter fed BLDC motor drive for solar PV array based water pumping," in *2014 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES)*, Mumbai, 2014.

[12] M. Banagar and A. Usha, "Design and simulation of bridgeless PFC buck-boost converter fed BLDC motor drive," in *2017 Innovations in Power and Advanced Computing Technologies (i-PACT)*, Vellore, 2017.

[13] C.-H. Wu and Y.-Y. Tzou, "Digital control strategy for efficiency optimization of a BLDC motor driver with VOPFC," in *2009 IEEE Energy Conversion Congress and Exposition*, San Jose, CA, 2009.

[14] V. Bist and B. Singh, "PFC Cuk Converter-Fed BLDC Motor Drive," *IEEE Transactions on Power Electronics,* vol. 30, no. 2, pp. 871-887, 2015.

[15] R. Silva-Ortigoza, V. M. Hernández-Guzmán, M. Antonio-Cruz and D. Muñoz-Carrillo, "DC/DC Buck Power Converter as a Smooth Starter for a DC Motor Based on a Hierarchical Control," *IEEE Transactions on Power Electronics,* vol. 30, no. 2, pp. 1076-1084, 2015.

[16] Z. Wu, Z. Genfu and J. Ying, "Line adaptive PAM & PWM drive for BLDCM," in *4th International Power Electronics and Motion Control Conference*, Xi'an, 2004.

[17] Y. S. Lai, K. Y. Lee, J. H. Tseng, Y. C. Chen and T. L. Hsiao, "Efficiency Comparison of PWM-Controlled and PAM-Controlled Sensorless BLDCM Drives for Refrigerator Applications," in *2007 IEEE Industry Applications Annual Meeting*, New Orleans, LA, 2007.

[18] M. M. J. Al-Ani, J. Carter and M. L. Jupp, "Comparison of electromagnetic performance and power losses of a high-speed machine fed by PWM and PAM inverter strategies," in *8th IET International Conference on Power Electronics, Machines, and Drives (PEMD 2016)*, Glasgow, 2016.

[19] R. Krishnan, Permanent Magnet Synchronous and Brushless DC Motor Drives, CRC Press, 2009.

[20] W. Huazhang, "Design and Implementation of Brushless DC Motor Drive and Control System," *Procedia Engineering,* vol. 29, pp. 2219-2224, 2012.

Signal Measurement Consistency of ECVT Data Acquisition System using Capacitor Array

Arbai Yusuf[1,2], Dodi Sudiana[1], Agus Santoso Tamsir[1], Melinda Melinda[1,3], and Harry Sudibyo S.[1]

[1]Department of Electrical Engineering, Universitas Indonesia, Depok, Indonesia
[2]CTECH Labs Edwar Technology Co, Tangerang, Indonesia
[3]Electrical Department, Engineering Faculty, University of Syiah Kuala, Banda Aceh, Indonesia
Email: arbai.yusuf41@ui.ac.id; dodi.sudiana@ui.ac.id; tamsir@ee.ui.ac.id; melinda@unsyiah.ac.id; harisudi@eng.ui.ac.id

***Abstract*—**This research relates to tomography based on capacitance measurement namely Electrical Capacitance Volume Tomography (ECVT). The paper discusses the design of ECVT data acquisition system that has frequent problems in measurement, such as error in signal measurement and fabrication errors. Based on those issue, development of an equipment that can help check the data acquisition system by using dummy capacitor array is proposed. In signal measurement, there are two modes: first without using array capacitor and second using capacitor array. By using a simple algorithm that refers to the measurement strategy, it is possible to immediately know which channel pair is experiencing problem. Measurement consistency on each channel is analyzed using the correlation coefficient (R) resulting an average value of 0.9387.

Keywords—ECVT; capacitive sensor; data acquisition system; charge amplifier circuit; capacitor array.

I. INTRODUCTION

Electrical Capacitance Volume Tomography (ECVT) is one of tomography based on capacitance measurement using capacitive sensor. The principle of capacitance tomography as shown in Fig. 1 is based on injection of square wave or sine wave signal to the electrode pair resulted in fringing electric field inside the sensor [1]. This signal is detected by other electrodes and controlled by an electronic circuit to be processed into an image using a specific reconstruction algorithm.

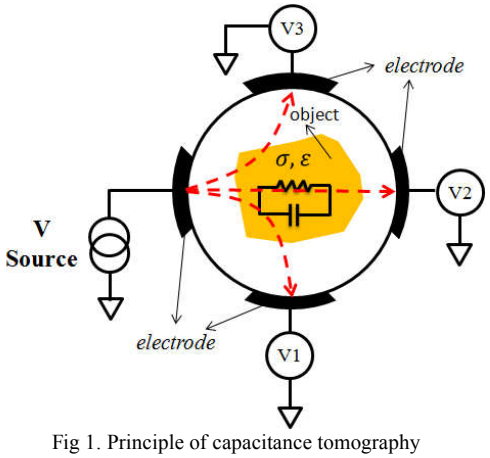


Fig 1. Principle of capacitance tomography

Currently, the ECVT data acquisition system has been developed accommodating 12, 16, 24, and up to 32 channels, in which each channel shall measure the capacitance independently [2]–[4]. Experiment that utilizes measurement from capacitive sensor often find some problems, i.e. missing data from ADC, output signal of charge amplifier are not present (too small) or even saturating circuit. Several researchers have attempted to improve the accuracy and reliability of the data acquisition system by optimizing the geometry of the capacitive sensor [5], adjusting the electrode configuration [6], reducing the effect of stray capacitance on analog switches [7], as well as quadrature phase detection [8].

According to the above issues, this paper proposes an equipment that can help in checking the data acquisition system by using dummy capacitor array. The array capacitor is made for eight electrode to check the measurement accuracy. Observation divided into four stage: first stage for first level, and then the array capacitor is moved to the next eight electrode at the second level, and so on up to the fourth level. Simple algorithm is employed to analyze the measurement in data analysis software to determine the signal consistency for all channels.

II. DESIGN

The ECVT data acquisition system has several main parts. The front end circuit contains analog switch configuration and charge amplifier circuit. Further stages are multiplexer circuit and subsequent digital signal processing circuit. Array of capacitors are designed in series and connected to the data acquisition system as shown in Fig. 2.

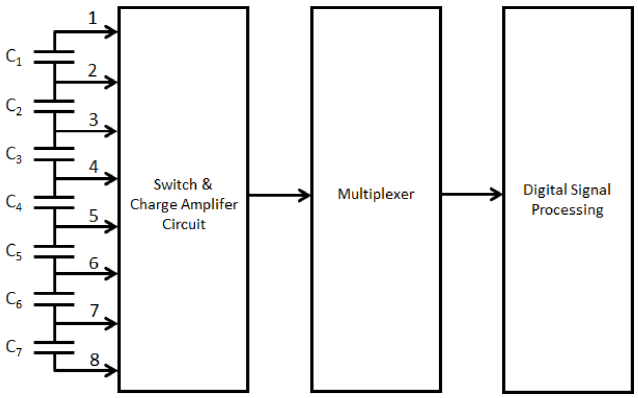


Fig 2. Design of equipment test using dummy capacitor array

As described above, the data acquisition system has maximum 32 channels, so as to test the system, four stages are required, with each stage consists of 8 channels. The system is tested in forward and reversed mode, hence 28 measurement data will be obtained as shown in Table 1.

TABLE I. MEASUREMENT STRATEGY USING FORWARD AND REVERSED MODE

No	Data	Forward	Reversed
1	V_1	1-2	2-1
2	V_2	2-3	3-2
3	V_3	3-4	4-3
4	V_4	4-5	5-4
5	V_5	5-6	6-5
6	V_6	6-7	7-6
7	V_7	7-8	8-7
8	V_8	9-10	10-9
9	V_9	10-11	11-10
10	V_{10}	11-12	12-11
11	V_{11}	12-13	13-12
12	V_{12}	13-14	14-13
13	V_{13}	14-15	15-14
14	V_{14}	15-16	16-15
15	V_{15}	17-18	18-17
16	V_{16}	18-19	19-18
17	V_{17}	19-20	20-19
18	V_{18}	20-21	21-20
19	V_{19}	21-22	22-21
20	V_{20}	22-23	23-22
21	V_{21}	23-24	24-23
22	V_{22}	25-26	26-25
23	V_{23}	26-27	27-26
24	V_{24}	27-28	28-27
25	V_{25}	28-29	29-28
26	V_{26}	29-30	30-29
27	V_{27}	30-31	31-30
28	V_{28}	31-32	32-31

III. SIGNAL MEASUREMENT

The basic principle of measuring capacitance in ECVT is using charge amplifier circuit as shown in Fig. 3, which consists of an Operational Amplifier (Op-Amp) and feedback components such as resistor R_f and capacitor C_f. $V_i(t)$ and $V_o(t)$ are sine wave signal input and output respectively, C_x is object capacitance measured in the sensor, whereas C_{s1} and C_{s2} are stray capacitances which are also measurable in the system. Stray capacitance is parasitic signal derived from other source such as from coaxial cable, screen sensor, and the effect of analog switches used to selecting electrode pair [9].

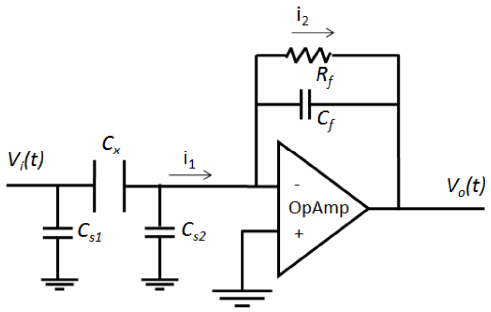


Fig 3. Charge amplifier circuit

Kirchhoff's law state that the current flowing to a point and the current flowing away from the point is the same $(i_1 = i_2)$, so that the output signal from this circuit is known. There are two methods of measuring capacitance, first by measuring the amplitude signal at the Op-Amp output, second, by measuring the phase difference between the excitation and the detection signal. The measurement capacitance utilizing the amplitude signal as described in this paper is shown in Eq. (1) and (2):

$$V_o = -k\frac{C_s}{C_f}V_i \quad (1)$$

$$k = \frac{R_f}{R_f + {}^{1}/_{\omega C_f}} \quad (2)$$

where V_i is excitation signal input, V_o is signal output, R_f is feedback resistance, C_f is feedback capacitance, C_s is capacitance to be measured inside sensor, and ω is angular frequency [10].

IV. RESULTS AND ANALISYS

There are two modes in signal measurement. First, initial measurement without capacitor array to find out how much stray capacitance in the data acquisition system. Second, measurement with dummy capacitor array to test the capability of measuring capacitive sensor appropriately. Experimental setup of ECVT data acquisition system using a dummy capacitor array is shown in Figure 4. In the following experiment, ECVT data acquisition system acquires data up to 32 channels which are divided into four levels with each level containing 8 channels. The capacitive sensor model uses dummy capacitor array with a fixed value of 1 pF.

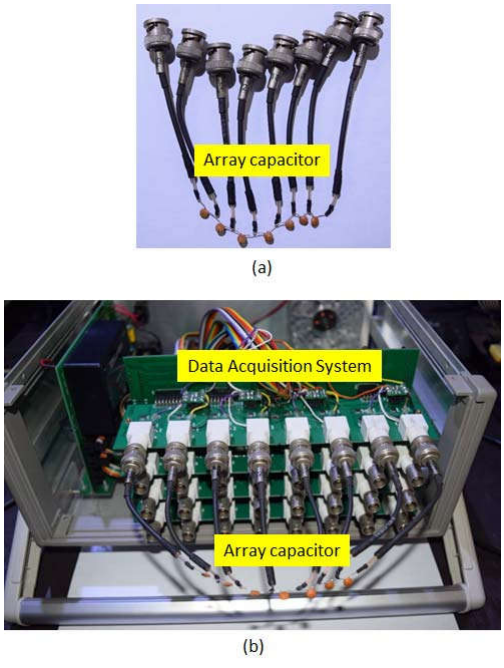


Fig 4. Measurement using dummy capacitor array

A. Measuring signal without dummy capacitor array

The input of data acquisition system on floating (not connected) is described as signal measurement without capacitor array. The purpose of the measurement with this mode is to find out how much stray capacitance in the data acquisition system. Stray capacitance constantly presents in the measurement system due to the influence of several factors such as the coaxial cable, the analog switch, and also signal interference from the electronic circuit. Measurements are done for each channel from channel 1 to channel 32 either forward or reversed. The forward measurement means the excitation voltage is distributed to the lower channel of each active electrode pair, while the other channel act as a detection. Examples are 1-2, 2-3, 3-4, 4-5, 5-6, etc. The reverse measurement means the voltage polarity is reversed, the excitation voltage is distributed to the upper channel of each active electrode pair, while the other channel act as a detection. Examples are 2-1, 3-2, 4-3, 5-4, 6-5, etc.

The measurement results are shown in Fig. 5 through 8, where each figure has two graphs: forward and reversed measurement. The 32 channel data acquisition system being used has four levels, each level contain eight channels. The first stage is measured at level one to obtain 7 data, then next level until fourth level so that 28 data measurement are obtained. A good system will produce measurable voltages and capacitances that have similar values, without significant signal spikes.

Figure 5 and 6 illustrate an example of inconsistent measurement with error in some channels. Figure 5 shows the output voltage of the charge amplifier circuit while Fig. 6 is the voltage data that has been converted to capacitance. According to the graph of measurement at first level (first 8 channel), good performance occurs where there is no error for either forward or reversed due to average signal uniformity in both voltage and capacitance. However, there are signal spikes on some channels. Measurement errors occur on the forward mode at level 2 on channel pair 10-11 and 12-13. While in reversed mode, there are errors at levels 3 and 4 for almost all channels due to the signal amplitude is greater than the normal level.

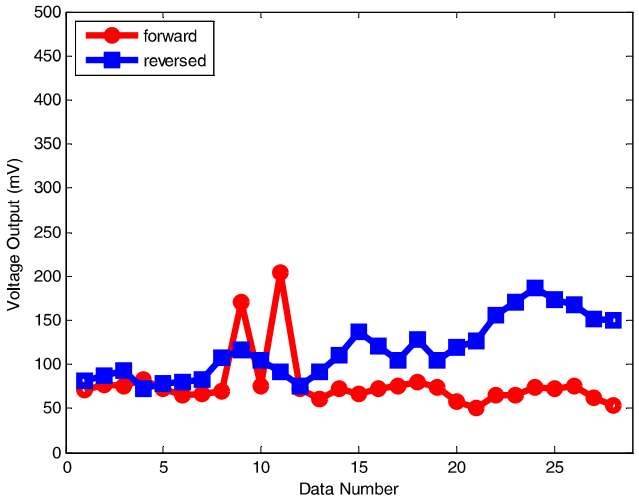


Fig 5. Inconsistency measurement of voltage output

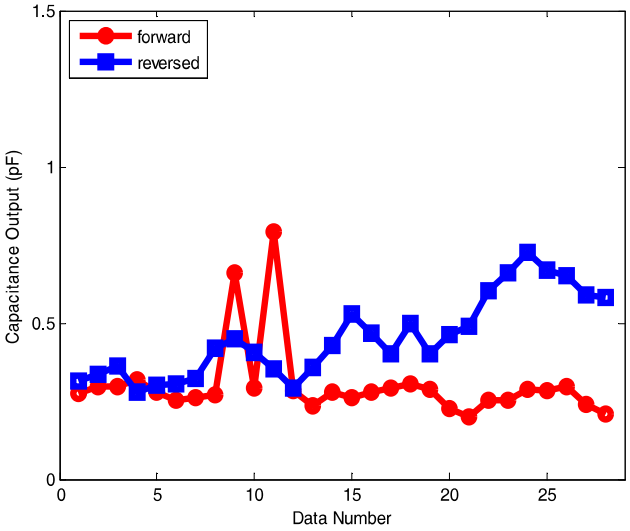


Fig 6. Voltage data converted to capacitance

Figure 7 and 8 show consistent measurement for all channel pairs with no error. In normal circumstances, the average voltage level will show 60-75 mV for both forward and reversed, or 250 fF (femto farad) after conversion into capacitance.

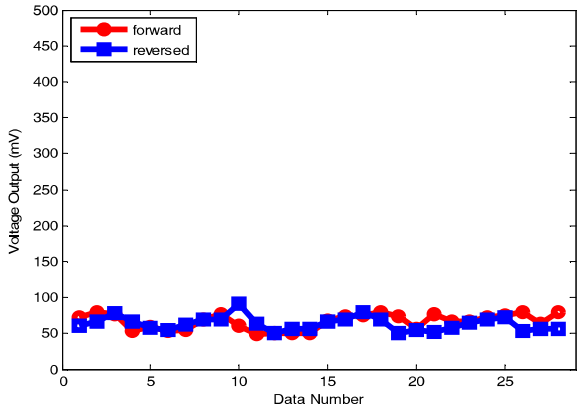


Fig 7. Consistency measurement of voltage output

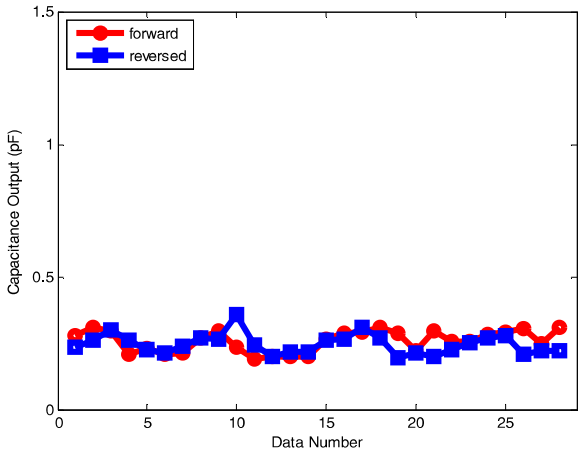


Fig 8. Voltage data converted to capacitance

B. *Measuring signal using dummy capacitor array*

Measurements using dummy capacitors array are performed to find out whether the ECVT data acquisition system is capable of measuring capacitive sensor appropriately. In this study, dummy capacitor array of 1pf is used. Measurements are done for each channel in both forward and reversed modes. The results of the measurements are shown in Fig. 9 to 12, where each figure has two graphs: forward and reversed measurement.

Figure 9 and 10 illustrate an example of inconsistent measurement with error in some channels. Figure 9 shows the output voltage of the charge amplifier circuit while Fig. 10 is the voltage data that has been converted to capacitance. According to the graph, the measurement at first level, second level, and third level are in agreement and there is no error either forward or reversed due the average signal uniformity for both voltage and capacitance. However, there are signal spikes on some channels. Measurement errors occur on the forward mode at level 4 on channel pair 28-29, 30-31 and 31-32. While in reversed mode, there are errors at level 4 on channel pair 29-28, 30-29, and 32-31. The result of capacitance measurement should approach dummy capacitor value about 1 pF, but since there are errors in some channels, it shows a considerably higher capacitance value of about 7 pF.

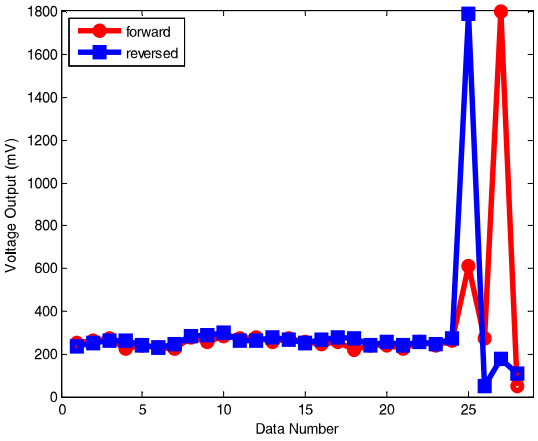


Fig 9. Inconsistency measurement of voltage output

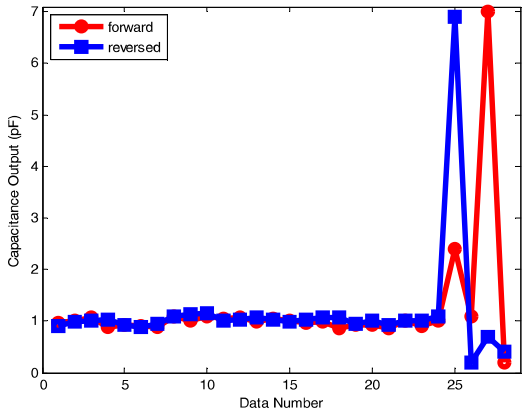


Fig 10. Voltage data converted to capacitance

Figure 11 and 12 show consistent measurement for all channel pairs with no error. In normal circumstances, the average voltage level will show 250 mV for both forward and reversed. The average value of capacitance measured by the system is about 1 pF for both the forward and reversed modes.

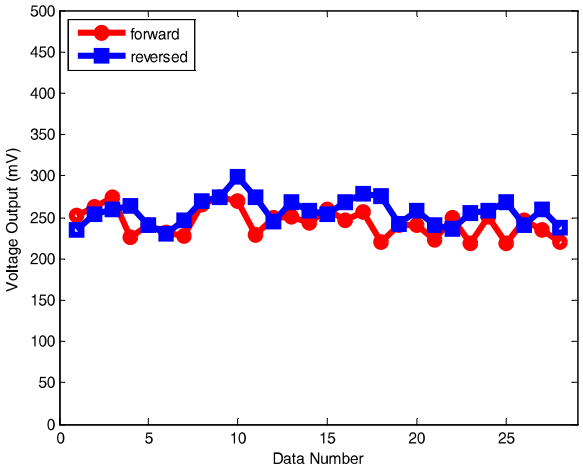


Fig 11. Consistency measurement of voltage output

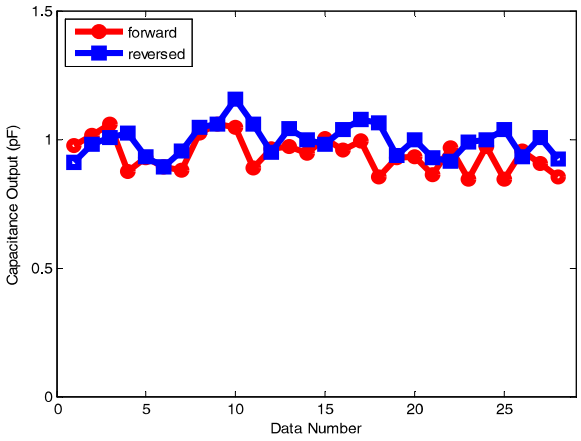


Fig 12. Voltage data converted to capacitance

C. *Measurement consistency*

Based on measurement signal of ECVT data acquisition system using dummy capacitor, the consistency of each channel is examined. The analysis method is coefficient correlation (R) by comparing the measurement results obtained by the system with dummy capacitor. The results of coefficient correlation are shown in Fig. 13 illustrating graphs in both forward and reversed modes. According to the graph, the value of coefficient correlation is close to 1 with an average of 0.9387. This indicates that the ECVT data acquisition system is reliable for use in practical applications, such as measuring the capacitive sensor.

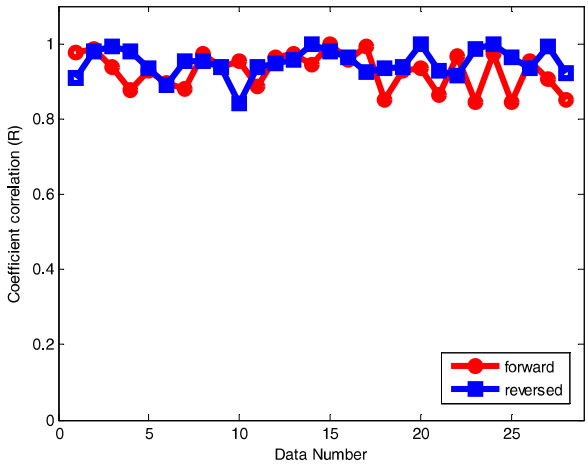


Fig 13. Coefficient Correlation (R) for dummy capacitor array measurement

V. Conclusion

ECVT data acquisition system has important requirement to ensure the uniformity of output signal. Signal spikes on some channels indicate error on specific channel pair, so that the measurement channel should be mitigated. This method has been tested using ECVT 32 channel data acquisition system. By using a simple algorithm that refers to the measurement strategy in Table 1, it is possible to immediately know which channel pair is experiencing problem. Measurement consistency of the system can be analyzed using correlation to assess the reliability of the system for practical application.

Acknowledgment

The author would like to thank all parties who kindly contribute to this research: Warsito P. Taruno, and Wahyu Widada from CTECH Labs Edwar Technology. This work is supported by a grant from Hibah PITTA DRPM Universitas Indonesia Grant Number: 723/UN2.R3.1/HKP.05.00/2017.

References

[1] D. Sudiana, A. S. Tamsir, and H. Sudibyo, "A digital signal processing algorithm on read out circuit for electrical capacitance tomography," in *Region 10 Conference (TENCON), 2016 IEEE*, 2016, pp. 1166–1170.

[2] A. Yusuf, I. Muttakin, A. Rudin, W. Widada, and W. P. Taruno, "Single signal conditioning multi electrode for ECVT data acquisition system," in *TENCON 2014-2014 IEEE Region 10 Conference*, 2014, pp. 1–6.

[3] K. Krzewska *et al.*, "Electrical Capacitance Volume Tomograph EVT4," in *7th World Congress on Industrial Process Tomography*, 2013.

[4] Y. Zhou, D. Liu, and H. Yan, "Three-dimensional electrical capacitance tomography measurement system based on an impedance analyzer," in *Imaging Systems and Techniques (IST), 2013 IEEE International Conference on*, 2013, pp. 415–418.

[5] Y. Li and D. J. Holland, "Optimizing the geometry of three-dimensional electrical capacitance tomography sensors," *IEEE Sens. J.*, vol. 15, no. 3, pp. 1567–1574, 2015.

[6] Y. Yang and L. Peng, "A configurable electrical capacitance tomography system using a combining electrode strategy," *Meas. Sci. Technol.*, vol. 24, no. 7, p. 74005, 2013.

[7] H. Sudibyo, D. Sudiana, A. S. Tamsir, I. Muttakin, W. Widada, and W. P. Taruno, "Switch Configuration Effect on Stray Capacitance in Electrical Capacitance Volume Tomography Hardware," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 14, no. 2, pp. 456–463, 2016.

[8] I. Muttakin, A. Yusuf, R. Rohmadi, W. Widada, and W. P. Taruno, "Design and simulation of quadrature phase detection in electrical capacitance volume tomography," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 13, no. 1, pp. 55–64, 2015.

[9] W. Q. Yang, "Hardware design of electrical capacitance tomography systems," *Meas. Sci. Technol.*, vol. 7, no. 3, p. 225, 1996.

[10] A. Yusuf, I. Muttakin, W. Widada, and W. P. Taruno, "Analysis of single excitation signal for high speed ECVT data acquisition system," in *Information Technology and Electrical Engineering (ICITEE), 2014 6th International Conference on*, 2014, pp. 1–6.

Control System Scheme for Shuttlecock Thrower Robotic Based on PID-CSA

Indra Ferdiansyah
Dept. Electrical Engineering
Politeknik Elektronika Negeri Surabaya Surabaya, Indonesia.
indraferdi@pens.ac.id

Bayu Sandi Martha
Dept. Informatics &Computer Engineering
Politeknik Elektronika Negeri Surabaya Surabaya, Indonesia.
bayu@pens.ac.id

Fernando Ardila
Dept. Informatics & Computer Engineering
Politeknik Elektronika Negeri Surabaya Surabaya, Indonesia.
nando@pens.ac.id

Haikal Firdiawan Zaky
Dept. Electrical Engineering
Politeknik Elektronika Negeri Surabaya Surabaya, Indonesia.
haikalzaky@ee.student.pens.ac.id

***Abstract*—The robot manipulator ability in maintaining its acceleration is very important due to its duties for having physical task such as moving the object and throwing the object. However the disturbance from the external system can be influenced to the accuracy which can reduce the robotic performance. Therefore the accuracy in doing the task should be constantly stable. Adding control mechanishm for keeping its acceleration is the solution for improving the robotic performance. In this paper, we propose control system for shuttlecock thrower robotic which is adopted for ABU ROBOCON 2018 themes. This control system is used for maintaining the acceleration and the elevationin throwing process can reach the specified target. Mechanical modification as well as the weight of object being thrown is also can be affected to the system. The proposed system is using Proportional, Integral and Derivative (PID) based on cuckoo search algorithm (CSA). The function of PID control is keeping the acceleration constantly in specified elevation. While the CSA can simplify the tunning PID control when it has some parameter modification. The result of this paper is robust control system from the disturbance which is applied in shuttlecock thrower robotic based its manipulation.**

Keywords—manipulator robotics, accuracy, shuttlecock throwing, PID, CSA

I. INTRODUCTION

Robot is one of mechanicals tool which can be used for doing many assignments as well as controlling requirement from human or defined programme. According to the construction, robotic has several types namely manipulator robotics, mobile robotics, humanoid robotics, and etc. In the robotics field espescially manipulator robotics, manipulation scheme for throwing the object is one of interesting topic in the robotics area[1, 2]. Accuracy is the important parameter at robotic design mainly for manipulator robotic as shuttlecock thrower should be keept its acceleration and elevation in stable condition [3, 4]. Figure 1 shows the application of robots in throwing an object to achieve a specified target. In accordance with the theme of ABU ROBOCON 2018, the shuttlecock (throwing ball) is thrown towards the “Golden Ring”, if after the ball goes through the golden-ring, then falling right into the "Golden Cup", the team manages to reach "Rong Bay" and is declared the winner.

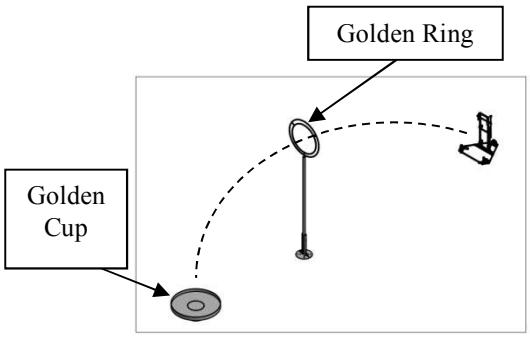


Fig. 1. Shuttlecock throwing manipulation on theme ABU ROBOCON 2018

At the manipulation process of the thrower, the robotic motion is dynamic required high level accuracy for getting the stable result [1],[5]. The type of actuator used is DC motor. Therefore this system is needed some mechanishm for holding the acceleration with control system.

There are various method which was used before for improving the performance of manipulator robotics. The main function is to make better performance of robotic especially robust from all kinds of disturbance. The dynamic load disturbance at the robotic can reduce its performance according to its low level accuracy[6,7,8].

In this paper, we propose the combination from PID control and CSA scheme to apply in this manipulator robotic. The purpose of PID control is protecting the actuator acceleration while throwing an object can appropriate to the specified elevation. The main step in PID control system is determining the constanta value which can be affected to the robotic performance. CSA is kind of methodology for optimizing the constanta value definition from PID control when there is load thrown modification, then the constanta value can be easily obtained which is called tunning process.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

II. THROWING MODELING

The modelling sytem of this manipulator robotic is based on throwing analysis from parabolic motion. According to the parabolic motion, there are two parameters for calculating the elevation and initial velocity. Those are maximum height (h_{max}) and the throwing distance (x_{max}) as defined to this following equation.

$$h_{max} = \frac{v_o^2 sin^2\alpha}{2\,g} \quad (1)$$

$$x_{max} = \frac{v_o^2 sin^2\alpha}{g} \quad (2)$$

According to Fig.2, manipulator robotic should be thrown the object to the suitable target. The throwing height and throwing distance have been determined before using equation (1) (2) for getting the elevation degree(α) and initial velocity (v_o) based on its appropriate target.

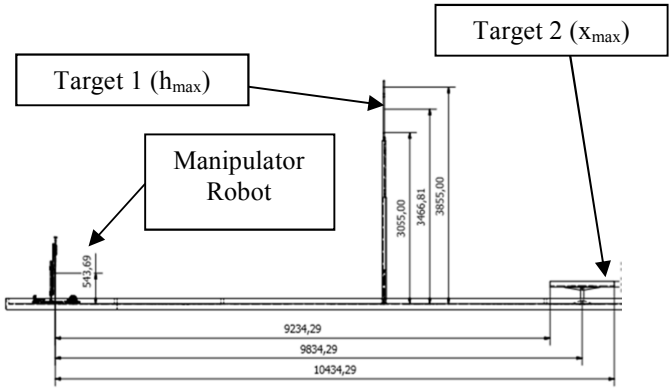


Fig. 2. Target "Shuttlecock Throwing" on field ABU Robocon 2018

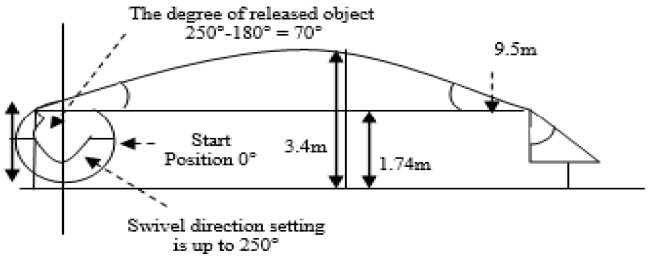


Fig. 3. Modeling manipulator robot for "Shuttlecock Throwing"

The throwing model of manipulator robotic is ilustrated at Fig.3., which have h_1 = 3.4metersand h_2=1.74meters, then the maximum height h_{max}= h_1-h_2 at maximum throwing distance x_{max}=9.5meters from the object position will be released using initial velocity v_o = 10.15m/s at h_2 height value until the object fall down to the last position. The distribution point of the falling object at the 0.4 meters height can be calculated using this following equation:

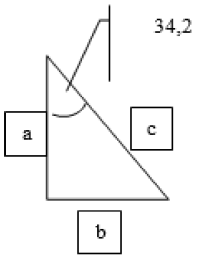


Fig. 4. The distribution model of the falling object at the last

$$a = h_2 - 0.4m \quad (3)$$

$$b = \frac{a\ sin\ 34.2}{\sin 55.8} \quad (4)$$

The distance of object to the last position in falling down can be calculated using this following equation:

$$x_{end} = x_{max} + b \quad (5)$$

Therefore using parabolic motion equation, the acceleration of the throwing process can be hold constantly with keeping the initial velocity value and calculating the elevation degree.

The manipulator robotic design is shown at Fig. 5. There is DC motor as the robotic activator in throwing process. Rotary encoder is also installed as the sensor detector of motor velocity and the elevation degree which will be used for control system input.



Fig. 5. Design manipulator robot for "Shuttlecock Throwing"

III. CONTROL METHOD

One of the most commonly used the control types at industry application is PID control. Designing simple control can be repaired the error system between process variable and set point value. The solution is calculating the error with some correcting process. The general equation of the PID control is described to this following equation:

$$u(t) = K_P \left(e(t) + \frac{1}{T_i}\int e(t)dt + T_d \frac{de(t)}{dt}\right) \quad (6)$$

Another equation can be formulated as:

$$u(t) = K_P\, e(t) + \frac{K_P}{T_i}\int e(t)dt + K_p\, T_d \frac{de(t)}{dt} \quad (7)$$

$$u(t) = K_P\, e(t) + K_i \int e(t)dt + K_d \frac{de(t)}{dt} \quad (8)$$

$$\text{If } K_i = \frac{K_P}{T_i} \text{ and } K_d = K_p\, T_d \quad (9)$$

Some parameters which will be used for PID control calculation are proportional value (Kp), Integral value (Ki) and derivative value (Kd). Those parameters can be influenced to the PID control responses. Hence, the value number parameter determination is the important factors for getting the best control result [9, 10].

Several methods can be used in this application, however CSA is the compatible method for this system. CSA methos adopt the cuckoo bird reproduction which have parasite characteristics and exploiting the other bird nest for incubating their eggs. Every cuckoo puts each egg at the certain time and throw their eggs to the randomly nest. The best nest with high quality of their eggs will be brought to the next generation. The number of available nest is constant and their eggs will be founded by the host bird at its nest with 0 and 1 probability[11, 12].

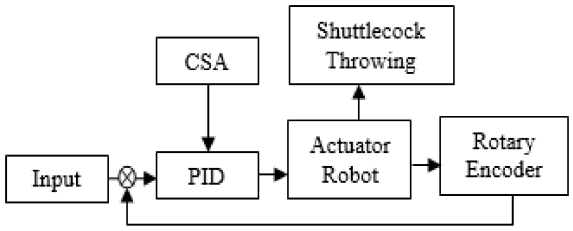


Fig. 6. Blok diagram tuning of PID-CSA

PID tunning process based on CSA method is ilustrated at Fig.6. The parameters value and flow diagram of this system are described to this following explanation:

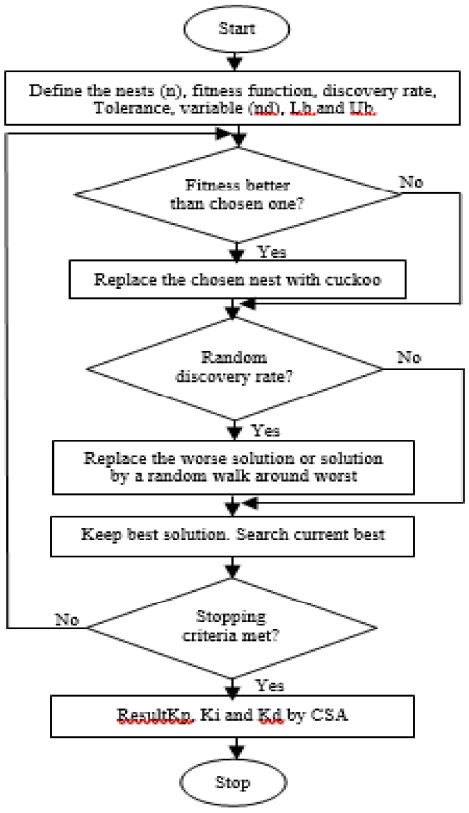


Fig. 7. CSA Flowchart

TABLE I. PARAMETERS OF CSA

Parameter	Value
Number of Nests (n)	25
Discover rate (pa)	0.25
Tolerance (Tol)	1.0e-5
Variable (nd)	15
Lb (Kp, Ki, Kd)	[3; 27.5; 0.03]
Ub (Kp, Ki, Kd)	[1.5; 22.5; 0.01]

IV. RESULT AND DISCUSSION

A. *Shuttlecoock Throwing Robot without Control*

There are two experiment of throwing process in this paper. First scheme is using open loop conditon without velocity controlling process and using specified elevation degree at 250°.

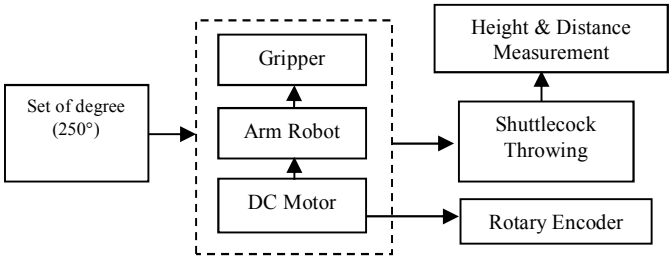


Fig. 8. Block of diagram Open Loop Shuttlecook Throwing

According to the Fig 8, the data measument is conducted in 20 times. The object which will be represented by shuttlecock will be clamped at robotics gripper. The motor DC as the actuator will be used for throwing process at open loop condition. Motor will be automaticaly spinned round after adding some energy. While at elevation degree250° has been formed, the gripper will be opened and the clamped object will be thrown. Data experiment of this measurement are height and the throwing distance which will be compared to the setpoint value. In this process, the error value number at throwing process can be obtained.

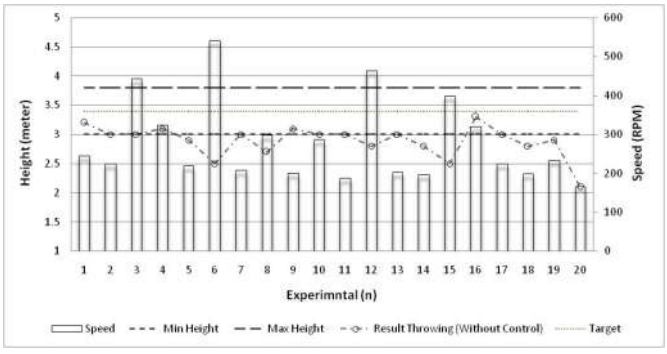


Fig. 9. Result of (height) throwing robot without control

The height of shuttlecock based on throwing process is shown at Fig.9. The expected target height is 3.4 meters which have some limitation in 3 meters up to 3.8 meters. When the height of throwing process is in that spesified range, the shuttlecock can through the target based on ABU robocon 2018 rules (TZ)

3. The graph at Fig. 10 show that according to the 20 times experiment of throwing process, there are only 4 data which are apropriate to the height range. The detailed experiment result are listed at Table II.

TABLE II. GOOD RESULT OF THROWING HEIGT WITHOUT CONTROL

No	Experimental (n)	Result
1	1	3.2 m
2	4	3.1 m
3	9	3.1 m
4	16	3.3 m

The height result from the throwing process will be validated using comparison graph between the height result to the expected height as the setpoint value.

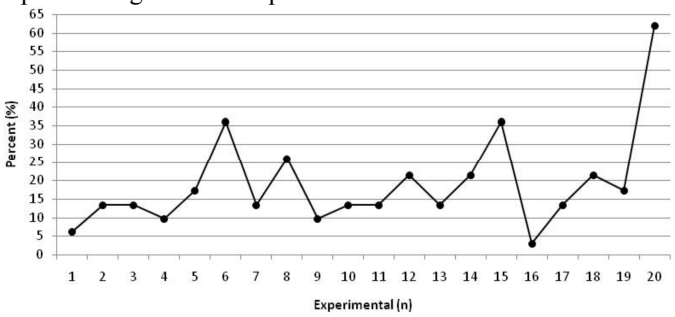


Fig. 10. Percentage error for (height) throwing robot without control

The error percentage value from each measurement are shown at Fig. 11. The smalest value is 3.03 % which have been occured at 16thexperiment. Due to the throwing process height value is almost same with the set point. While the throwing process is far away from the set point value, the error percentage can increase up to 61.90 % at 16thexperiment.In an experiment with errors below 10%, the shuttlecock can pass through the golden ring with an average percentage of shuttlecock throwing height error from the uncontrolled robot of 19.02%. Thus, to pass through the golden ring, a control system is required that is able to keep the throwing height has a percentage error below 10%.

The next data experiment is throwing distance data which is in the range of 9.2 meters up to 10.4 meters based on champion rules at TZ. In this system is using 9.8 meters distance as the setpoint value.

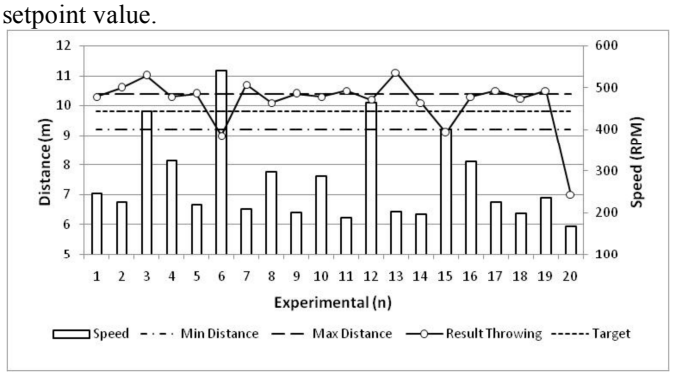


Fig. 11. Result of (distance) throwing robot without control

As well as height measurement, in distance measurement there are 20 measurement data which are seven measurements are in the throwing distance range. The result of this measurement are listed at Table III.

TABLE III. GOOD RESULT OF THROWING DISTANCE WITHOUT CONTROL

No	Experimental (n)	Result
1	4	10.3 m
2	8	10.1 m
3	10	10.3 m
4	12	10.2 m
5	14	10.1 m
6	16	10.3 m
7	18	10.25 m

The distance data result of this measurement is 10.2 meters based on its range of throwing distance. The detailed difference between the result data and expected data will be compared as ilustrated in Fig. 12.

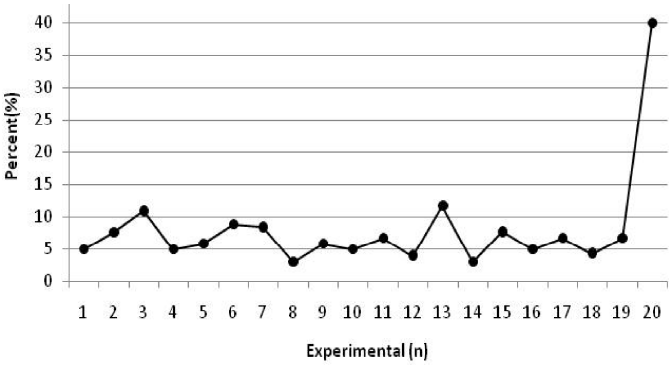


Fig. 12. Percentage error for (distance) throwing robot without control

The maximum error limitation of this error distance should be 5 % for achieving range position. The highest error is up to 40% due to the distance result is only 7 meters which is lower than 9 meters target. Another result show the error percentage is lower than 15 % and its average is only 8.013 %.Thus in order to achieve the golden cup, the control system must maintain the percentage of errors from the distance of throwing results below 5%.

According to the 20 numbers of experiment result, the experiment which can be reached the expected target of TZ 3 rules is only 4th and 16thexperiment. Those experiment have compatible throwing result based on distance and height range.

B. Shuttlecoock Throwing Robot using PID-CSA Control

Another throwing scheme of this paper is using control system to the robotic actuator. The function of this control is keeping the accceleration at the stable condition. The fixed value of initial velocity should be stayed at 250^0 degree elevataion.

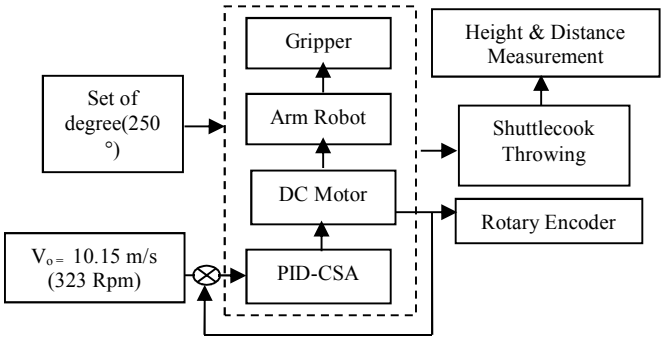


Fig. 13. Block of diagram Close Loop Shuttlecook Throwing

There is feedback system at DC motor for holding the velocity, in order to getting stable acceleration in throwing process. As well as previous experiment, the kind of data measurement are distance and height from throwing process of the robotic.

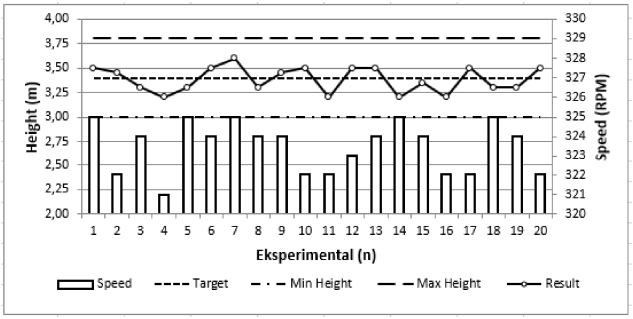


Fig. 14. Result of (height) throwing robot using PID-CSA control

Fig.14 show the result of robotic throwing process in height data measurement. There are twenty times data measurement which can be reached the 3.3 meters of height target. All of height data result are stable although there is small difference between the measurement and the setpoint value. The detailed analysis can be calculated using error percentage for each throwing.

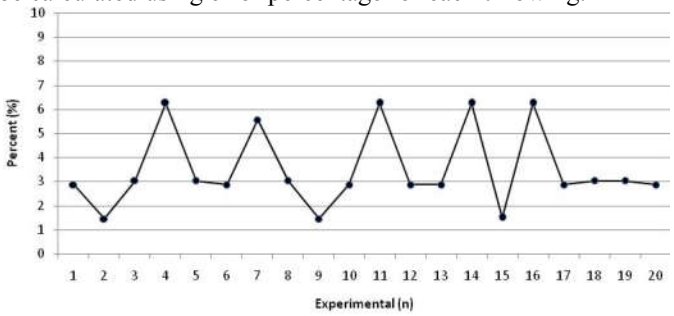


Fig. 15. Percentage error for (height) throwing robot using PID-CSA control

The highest value of error percentage at Fig.15 are occured at 4th, 11th, 14th and 16th experiment up to 6.25%. While at 2th, 9th and 15th experiment are only 1.4% which are relatively smaller than before. The average value of error percentage from throwing result using PID-CSA is 3.5%.For finishing all target at TZ rules, the expected distance should be in range of the rong bray.

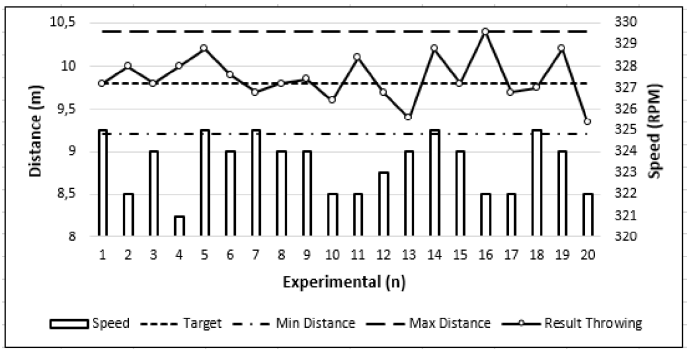


Fig. 16. Result of (distance) throwing robot using PID-CSA control

In this experiment, we analyze the distance result based on the throwing process using control system. The average distance result value is 9.82 meters which should be categorized into expected range. The result show that the value is approaching to the 9.8 meters as the expected value. While, there are some experimental result still quite far away from expected value such as at 5th, 14th and 19th experiment. Those experiment have difference value up to 13 meters.

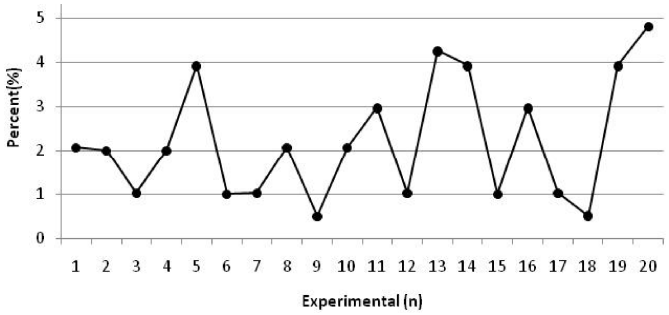


Fig. 17. Percentage error for (distance) throwing robot using PID-CSA control

Three largest error of this experimental result are 4.81 % at 20th experiment, 4.25% at 13th experiment and 3.92% at 5th experiment.Based on all experimental results, the proposed system can be declared stable because the shuttlecock reaches all targets ("Golden Ring" and "Golden Cup") and has an error percentage value below 5% with an average value of 2.20%.

The detail of the throwing process performed using the PID-CSA control is shown in Figure 18. From the figure it can be seen that in the first stage the object starts to be moved / swung (image order no 1) to reach the specified angle with the initial velocity (Vo) . Then the object begins to be released (image order 2, 3 and 4) moves towards the first target of TZ 3 to pass a pole with an altitude above 3 m (image order no 5). After able to pass the first target (image order no 6) then the object will go to the second target of rong bay box (image order no 7 and 8) which is more than 9.5 m from the position of robot. Finally, the object goes into the rong bay (imageorder no 9)



Fig. 18. Snapshots of the Shuttlecock throwing of into the rong bay

V. CONCLUSION

In this paper, we propose PID-CSA scheme for maintaining the acceleration stability. This proposed scheme is applied to the manipulator robot when throwing the object into the exactly hole. The result show that using PID-CSA can hold the velocity in a certain elevation degree. It proves from the error result of the height can be reduces up to 15.52% and 5.81% for the distance reduction. PID-CSA control system is better solution for maintaining the throwing process of manipulator robotic.

REFERENCES

[1] H. Miyashita, T. Yamawaki and M. Yashima, "Learning Control Method for Throwing an Object More Accurately with One Degree of Freedom Robot," *2010 IEEE/ASME International Conference on Advanced Intellegent Mechatronics,* July 2010,pp. 397-402.

[2] M. Buehler, D. E. Koditschek and P. J. Kindlmann, "Planning and Control of Robotic Juggling and Catching Tasks,"*International Journal of Robotics Research*, Vol.13, No.2, 1994, pp.101-118.

[3] H. Miyashita, T. Yamawaki and M. Yashima, "Control for Throwing Manipulation by One Joint Robot," *2009 IEEE International Conference on Robotics and Automation Kobe International Conference Centre,* May 2009,pp. 43-48.

[4] H. Miyashita, T. Yamawaki and M. Yashima, "Parts Assembly by Throwing Manipulation with a One-Joint Arm," *2010 IEEE/RJS International Conference on Intelligent Robots and Systems,*October 2010,pp. 1273-1278.

[5] M. T. Mason, "Mechanics of Robotic Manipulation, " The MIT Press 2001.

[6] V. Falkenhahn, T. Mahl, A. Hildebrandt, R. Neumann and O. Sawodny, "Dynamic modeling of constant curvature continuum robots using the Euler-Lagrange formalism", *International Conference on Intelligent Robots and Systems*, Chicago, 2014, pp. 2428-2433.

[7] P. Dong, P. Chanhun, K. Hwisu, P. Joohan and K. Doohyung, "Dynamic simulation of the robot manipulator for mobile based rescue robot", *The 12th International Conference on Ubiquitous Robots and Ambient Intelligence*, 2015, pp. 345-346.

[8] Z. Sameh, B. Wissem, T. Hatem, K. Jilani and P. Vicenc, "Comparison between ultra-local model and auto-tuning PID controls strategies applied to the redundant manipulator robot for several inputs", *18th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA),* 2017, pp. 350-355

[9] T.S. Jonathan, R. Levi and K.D Marc, " Comparison of linearized dynamic robot manipulator models for models predictive control", *IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids),* 2017, pp.205-212.

[10] G. Yanping, X. Zhang, H. Donggang and D. Chnaghui, "The Mobile Control Scheme of Intelligent Manipulator Based on the Digital PID," *2010 Second International Symposium of Intelligent Information Technology Application,* 2008, pp. 703-707.

[11] S. Jagindar, I. Elamvazuthi, Shaari and N. Perumal, "Development of PID Controller Tuning Tool based on Cuckoo Search Algorithms," *IEEE 3rd International Symposium on Robotics and Manufacturing Automation (ROMA),* 2017.

[12] S. Richa, G. Prerna and A. P. Mittal, "Performance evaluation of cuckoo search algorithm based FOPID controllers applied to a robotic manipulator with actuator,*"International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India,* 2015, pp. 356-36.

A Natural Phenomenon Emphasizing of the Thunderstorm Processes and Salmon Migration for the Optimal Power Production of the Power System Operation

A.N. Afandi[1], Irham Fadlika[2], Yuni Rahmawati[3], Quota Alief Sias[4], Langlang Gumilar[5], Irawan Dwi Wahyono[6]
Electrical Engineering, Universitas Negeri Malang
Malang, Jawa Timur, Indonesia
[1]an.afandi@um.ac.id, [2]irham.fadlika.ft@um.ac.id,
[3]yuni.rahmawati.ft@um.ac.id, [4]quota.alief.ft@um.ac.id,
[5]langlang.gumilar.ft@um.ac.id, [6]irawan.dwi.ft@um.ac.id

Yunis Sulistyorini
Mathematical Education, IKIP Budi Utomo
Malang, Jatim, Indonesia
yunis.sulistyorini@gmail.com

***Abstract*—This paper presents an adoption of a natural phenomenon as Thunderstorm Algorithm (TA) which is applied to solve a problem of the power production composition under various constraints. This work also introduces artificial salmon tracking algorithm (ASTA) for defining the optimal strategy of the power system on the power consumption. Both algorithms are tested on the IEEE-62 bus system as a selected structure for the mathematical cased model. By considering all parameters, results show that ASTA can be applied to predict the power consumption and TA also has good performances while searching the optimal solution. Moreover, the power production can be presented throughout an economic dispatch problem. Technically, this computation demonstrates the optimal solution with fast convergence and short time consumption. These processes also perform smooth and stable characteristics for the searching completion.**

Keywords—Economic dispatch; emission dispatch; operating cost; salmon tracking; thunderstorm algorithm.

I. INTRODUCTION

Power demand is one of the most factors which is used to be a main consideration on the power production. This factor is associated with energy usages on the consumer sites patterns in consumer behavior related to variations over the day, night, week, seasons and holidays. In general, the power commodity should be designed in a procurement balance between supply and consumption in the power market. Deal with this condition, a demand forecasting (DF) strategy is very important to meet the power production in the behavior of the energy players. In addition, the optimal power production is also urgent to cover all possible combination between existing power plants as a unit commitment. The unit commitment is commonly approached using an economic dispatch (ED) problem [1]–[8]. The ED is used to find the optimal allocation of power outputs among the various generators available to serve the load [6], [9]–[14]. Recently, the ED also considers an emission dispatch (EmD) problem for decreasing pollutants combined as an economic power system operation problem [1], [15]–[18].

Presently, various types of the intelligent computation (IC) are more popular than classical approaches which are applied to solve the optimized power system problem. These methods were formed to bring biological inspirations and natural entity behaviors become computational approaches. It has also shifted classical usages in many implementations on optimization problems [2], [13], [19]–[24]. Recently, the IC covered evolutionary algorithms has started to gain significant attention during implementations on wide areas with many operational constraints for replacing weaknesses of classical approaches. In principles, the synthesis and analysis of computational agents that act intelligently are major processes in the IC whose make decisions in terms of computing aspects. Moreover, the agents interact in complex ways and provide stages for the different learning associated with its natural inspiration.

Since the early idea of the IC, many methods have been proposed based on own inspirations [16], [25] and many natural phenomena or biological processes have been also adopted as the inspiration [15], [25], [26]. Currently, many algorithms have been proposed which are conducted to phenomena or entities in nature [2], [11], [26]–[29]. These versions have been advanced to increase computational performances [1], [9], [30]–[32]. In details, hierarchies and procedures are developed specifically to maintain the computational ability as given in sequencing orders of the algorithm. In line with previous efforts, this paper presents an emphasizing of thunderstorm mechanism and salmon tracking applied to the power system problem considered various operational limitations.

II. NATURAL PHENOMENON EMPHASIZING

Revealing natural phenomena or behaviors are subjected to emphasizing cognitive processes in nature. The advanced point of these processes has been also related to the learning reinforcement for understanding how behavioral programs control the selection procedures. In these works, the emphasis is going on the concepts of the natural phenomenon and behavior [25], [33]–[36].

In nature, a thunderstorm can be recognized by cloud shapes and a pre-signal, such as, like a puffy shape expanded to begin the growing upwards of the potential striking path between clouds and the earth. By considering the charge, a lightning process is ignited by the form of a spark in a charged cloud as the interaction between the negative charge deployed in the lower part and the positive charge center is located at the opposite up [35], [37], [38]. By considering a phenomenon of a thunderstorm, its mechanisms are adopted as an inspiration to explore a novel intelligent computation called in Thunderstorm Algorithm (TA) as illustrated in Fig. 1.

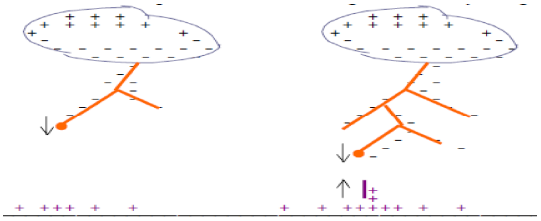


Fig. 1. Illustration of the striking propagation

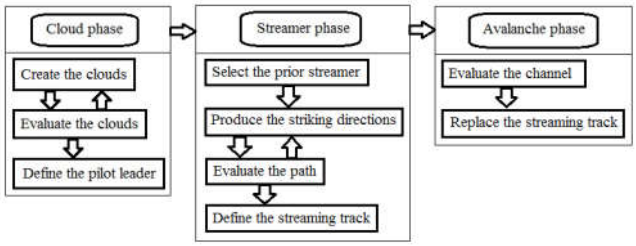


Fig. 2. Hierarchy processes of the thunderstorm algorithm

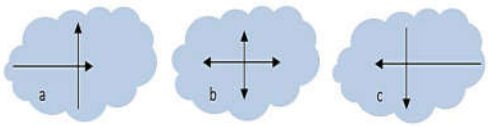


Fig. 3. Cloud's mechanisms (a. forward mechanism, b. expanded mechanism, and c. reverse mechanism)

As given in Fig. 2, the searching mechanism of TA is conducted to striking processes and channeling avalanches. In addition, the cloud charge is populated as illustrated in Fig. 3. This figure illustrates for the forward mechanism, expanded mechanism, and reverse mechanism. In principle, the sequencing computation of TA is given in pseudo-codes in terms of Cloud Phase: Streamer Phase; and Avalanche Phase [9], [12], [32].

In particular, salmon fish movements are characterized based on fresh water, feeding territories, and spawning movements. Moreover, its migration covers an adaptive specialization for downstream migration, seawater entry, and marine residence. By considering these steps, physiological changes include increased salinity tolerance, olfactory sensitivity, metabolic rate, the scope for growth, and altered hemoglobin and visual pigments. In general, salmon life history can be illustrated using natural behaviors in terms of spawning fish and returning from the ocean. In addition, Fig. 4 shows the lifetime of salmon cycles.

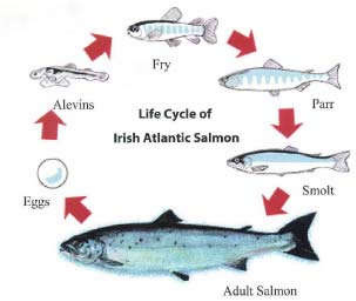


Fig. 4. Life cycle of salmon fish

Habitually, the salmon run is the moving time migrated from the ocean, swim to the upper reaches of rivers where spawning on gravel beds [39]. The early life of salmon spends in rivers continue to swim out to sea for the living adult. For many years ago, many studies were explored and developed in wide variety models to understand the migration situation, such as simple moving-average time series models, complex stock-recruitment models, and non-mechanistic models [33], [34], [36], [39]. By considering these behaviors, artificial salmon tracking algorithm (ASTA) is constructed based on migrating steps as a computational intelligence.

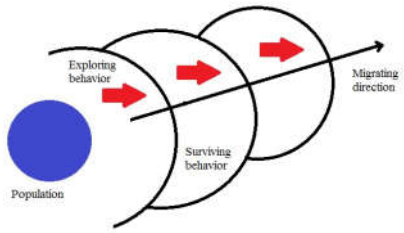


Fig. 5. Principles of the artificial salmon tracking algorithm

Technically, ASTA is compiled using its procedures referred to the exploring and surviving steps as illustrated in Fig. 5. The exploring step is used to search out a mouth river for guiding the desired possibility selection. The surviving step is used to find out the returning destination to track the desired solution at all various branches. Moreover, ASTA is presented in terms of salmon number, surviving factor, mouth river, tracking round, migrating period.

III. Engineering Model Reinforcement

Available models of many technical problems are used to fit the experimental data, simulation, and forecasting. Specifically, a processing model can be derived in certain variations and concentrations based on parameters to present the real situation and condition. In details, the model can also be also used to plan the amount of evaluation required to achieve desired levels of various fields [3]–[5], [7], [8], [15]. In this section, the model leads to the ED, EmD, and DF problems. The model is used to find the optimal allocation of power outputs among the various generators available to serve the load [9], [10], [10]–[12], [20], [40]–[42]. By the balanced procurement, these processes are conducted to the DF for

operating cost reduction while producing power outputs to meet a total demand.

The important things of the ED cover technical cost of products with the pollutant reduction on energy services. To solve the ED combined with the EmD, many strategies are developed to explore an energy function composing considered the fuel cost, emission, transmission line losses, weighting factors, and others [5], [23]–[25]. In these works, the ED also covers an environmental protection for reducing atmospheric emissions at thermal power plants approached using an emission dispatch (EmD) problem for decreasing pollutants [1], [15]–[17]. This emission dispatch is subjected to reduce various pollutant discharges [28], [43]. By considering ED and EmD problems, these studies are presented in single objective function as the economic and emission dispatch (EED) problem under operational limitations.

Furthermore, many important decisions are made by describing the EED problem for selecting some economical measurements on the desirable targets [1], [7], [40], [42], [44]. The EED has also included penalty and compromised factors. Moreover, a weighting factor is also often used to be a parameter for updating the generating unit contributor. In these studies, the dispatching problem is formulated using mathematical statements for defining the objective function and technical constraints [11], [17], [29], [43].

IV. Applied Algorithm Procedures

In general, many previous works have been addressed to several IEEE models. In these studies, the IEEE-62 bus system is selected as a sample model of the power system. In detail, TA and ASTA are designed based on its structures and procedures. Technically, this model consists of 62 buses; 89 lines; and 32 load buses. This system is also supported by 19 generating units with its coefficients and designed limits as given in Table I, Table II, Table III, and Table IV. In detail, these works are required by 10% of the loss limit; 0.5 of the weighting factor; and 0.85 kg/h of the emission standard; 5% of voltage violations; 95% of the power transfer capability; and banded on upper and lower power limits. Table 1 and Table 2 illustrate the quadratic coefficients for the fuel cost consumptions covered 19 generating units while emission coefficients of generating units are listed in Table III and Table IV with the power limits

TABLE I. Fuel Cost Coefficients of Cheaper Generators

a ($/MWh^2)	b ($/MWh)	c	Gen
0.00550	4.00	30	G2
0.00550	4.00	45	G3
0.00250	0.85	10	G4
0.00550	4.00	90	G6
0.00200	0.50	58	G10
0.00450	1.60	65	G11
0.00250	0.85	78	G12
0.00500	1.80	75	G13
0.00450	1.60	85	G14
0.00450	1.40	90	G16
0.00250	0.85	10	G17
0.00450	1.60	25	G18

TABLE II. Fuel Cost Coefficients of Expensive Generators

a ($/MWh^2)	b ($/MWh)	c	Gen
0.00700	6.80	95	G1
0.00600	4.60	20	G5
0.00650	4.70	42	G7
0.00750	5.00	46	G8
0.00850	6.00	55	G9
0.00650	4.70	80	G15
0.00800	5.50	90	G19

TABLE III. Low Emission Coefficients and Power Limits of Generators

α (kg/MWh^2)	β (kg/MWh)	γ	P_{min} (MW)	P_{max} (MW)	Gen
0.0180	-1.8100	24.300	50	300	G1
0.0136	-1.3000	22.070	0	100	G4
0.0180	-1.8100	24.300	50	300	G5
0.0126	-1.3600	23.040	50	200	G7
0.0136	-1.3000	22.070	0	100	G10
0.0139	-1.2500	23.010	50	150	G11
0.0121	-1.2700	21.090	0	100	G12
0.0180	-1.8100	24.300	50	300	G13
0.0140	-1.2000	23.060	0	150	G14
0.0139	-1.2500	23.010	50	150	G16
0.0136	-1.3000	22.070	0	100	G17
0.0180	-1.8100	24.300	50	300	G18

TABLE IV. High Emission Coefficients and Power Limits of Generators

α (kg/MWh^2)	β (kg/MWh)	γ	P_{min} (MW)	P_{max} (MW)	Gen
0.0330	-2.5000	27.023	50	450	G2
0.0330	-2.5000	27.023	50	450	G3
0.0330	-2.5000	27.023	50	450	G6
0.0360	-3.0000	29.030	50	500	G8
0.0400	-3.2000	27.050	0	600	G9
0.0360	-3.0000	29.000	0	500	G15
0.0400	-3.000	27.010	80	600	G19

In this section, TA is implemented on a standard model of the power system based on the sequencing orders as depicted in Fig. 2. In these works, TA is compiled using the cloud phase; streamer phase; and avalanche phase in terms of 1 of the avalanche; 25 of the cloud charge; 100 of the streaming flow; and 4 of the hazardous factor. On the other hand, ASTA is combined using pseudo-codes for searching the best solution covered for the salmon number, surviving factor, mouth river, tracking round, migrating period. In details, ASTA is presented using 100 of the salmon number, 0.25 of a surviving factor, 100 of the mouth river, 100 of the tracking round, 1 of the Migrating period, and 50 of the population.

V. Results and Discussions

In these works, the simulation is addressed to evaluate a computing ability while searching the optimal solution of the EED problem as the optimal power production. Based on the technical requirements, the power balance of joined generating units considers 2,221.2 MW and 1,055.8 MVar of the load

demand. In detail, the strategy of the production is decided by the optimal sharing portion for all contributors associated with low operating costs and pollutant limitations. The final results are given in following tables and figures.

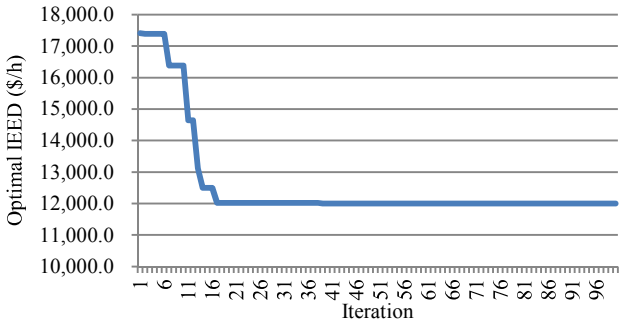


Fig. 6. Computational speed of the optimal solution

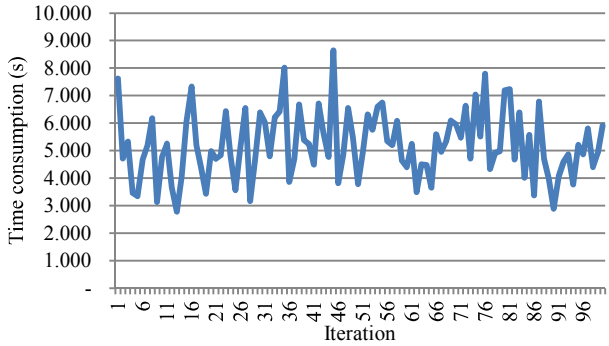


Fig. 7. Computational time consumption

TABLE V. POWER OUTPUT PRODUCTIONS

Gen	Power Limits (MW)			Power Output (MW)	Remarks
	P_{min} (MW)	P_{max} (MW)	P_{mid} (MW)		
G1	50	300	175	117.2	Low
G2	50	450	250	97.5	Low
G3	50	450	250	110.6	Low
G4	0	100	50	93.4	High
G5	50	300	175	115.6	Low
G6	50	450	250	167.5	Low
G7	50	200	125	102.6	Low
G8	50	500	275	115.0	Low
G9	0	600	300	122.6	Low
G10	0	100	50	75.3	High
G11	50	150	100	150.0	Middle
G12	0	100	50	98.7	High
G13	50	300	175	175.0	Middle
G14	0	150	75	142.5	High
G15	0	500	250	133.2	Low
G16	50	150	100	137.9	High
G17	0	100	50	93.2	High
G18	50	300	175	144.8	Low
G19	100	600	350	195.3	Low

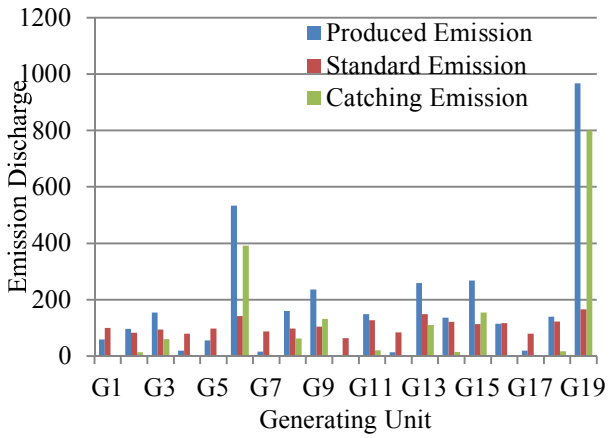


Fig. 8. Emission discharges of generating units

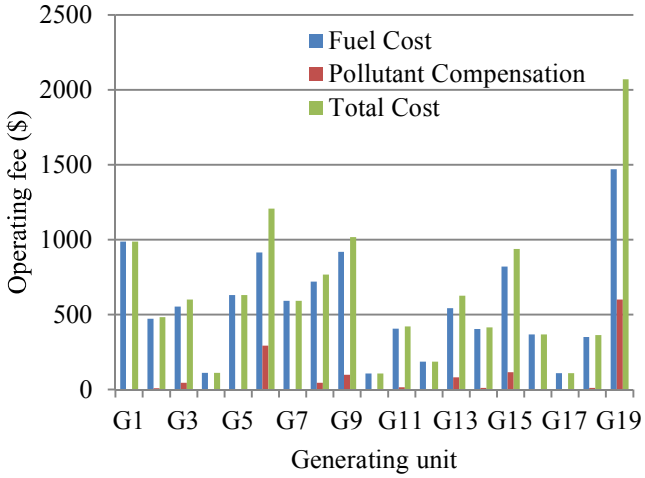


Fig. 9. Operating fees of generating units

From graphical performances, it is known that Fig. 6 initiates a computational convergence which is obtained in 17 iterations. This solution is also searched in 5.2 s for the charge of 12,017 $/h. This characteristic also describes smooth and stable processes. The time consumption of the computation is presented in Fig. 7 associated with different amount while searching for the solution for each iteration. In total, this simulation is completed in 519.5 s. For all executions, the computation results are covered in Table V for the power balance.

From Table V, the unit commitment of generating units is produced totally around 2,387.9 MW. It means that the power system has 166.7 MW of the loss during conveying energy to the load area around 2,221.2 MW. It can understand that the lines of the power system consume the power of around 6.9% of the power production. As the implication of this unit commitment, generating units also produce individually power outputs within various portions of the pollution. In total, the emission is produced around 3,398.6 kg/h even the pollution is also limited totally by 2,029.7 kg/h. The higher contributors

for the emission are released from G6; G15; and G19 as detailed in Figure 8. Caused by over productions, these emissions should be filtered at generating units around 1,778.4 kg/h. Furthermore, as impacts of the combination, the power system is operated using 12,005.6 $/h for the fuel procurement and the emission compensation. In Particular, the financial results are given in 10,671.8 $/h which is spent for the fuel consumption and 1,333.8 $/h is used to provide protectors or to pay the pollutant compensation. In details, individual operating fees are presented in Figure 9.

VI. Conclusions

This paper presents a natural learning based on an engineering reinforcement to carry out the economic and emission dispatch as a strategy to meet the optimal composition of generating units. Algorithms are applied to an IEEE-62 bus system by considering compositions of cost and emission aspects. Results obtained shows that the optimal solution is revealed in different individual power outputs, emission discharges, and optimized operating costs. From these works, the opportunity future studies are devoted to real power system applications and algorithm's parameter observations.

Acknowledgment

The authors gratefully acknowledge the support of Universitas Negeri Malang, Indonesia, for the PNBP Research Program issued on 21.2.6/UN32.14/LT/2018.

References

[1] A. N. Afandi, "Optimal scheduling power generations using HSABC algorithm considered a new penalty factor approach," in *The 2nd IEEE Conference on Power Engineering and Renewable Energy (ICPERE) 2014*, 2014, pp. 13–18.

[2] S. Chandra, V. Kothari, and M. Sharma, "A model for controlling variance in the Artificial Bee Colony algorithm," in *2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2015, pp. 263–269.

[3] A. N. Afandi, *Optimal Solution of the EPED Problem Considering Space Areas of HSABC on the Power System Operation*, vol. 7. 2015.

[4] N. Tutkun, O. Can, and A. N. Afandi, "Low cost operation of an off-grid wind-PV system electrifying residential homes through combinatorial optimization by the RCGA," in *2017 5th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, 2017, pp. 38–42.

[5] M. El-Shimy, M. A. Attia, N. Mostafa, and A. N. Afandi, "Performance of grid-connected wind power plants as affected by load models: A comparative study," in *2017 5th International Conference on Electrical, Electronics and Information Engineering (ICEEIE)*, 2017, pp. 1–8.

[6] A. N. Afandi and Y. Sulistyorini, "Thunderstorm Algorithm for Determining Unit Commitment in Power System Operation," *J. Eng. Technol. Sci.*, vol. 48, no. 6, pp. 743–752, Dec. 2016.

[7] Y. Li, S. Miao, X. Luo, and J. Wang, "Optimization model for the power system scheduling with wind generation and compressed air energy storage combination," in *2016 22nd International Conference on Automation and Computing (ICAC)*, 2016, pp. 300–305.

[8] A. N. Afandi, I. Fadlika, and A. Andoko, "Comparing Performances of Evolutionary Algorithms on the Emission Dispatch and Economic Dispatch Problem," *TELKOMNIKA Telecommun. Comput. Electron. Control*, vol. 13, no. 4, pp. 1187–1193, Dec. 2015.

[9] A. N. Afandi *et al.*, "Designed Operating Approach of Economic Dispatch for Java Bali Power Grid Areas Considered Wind Energy and Pollutant Emission Optimized Using Thunderstorm Algorithm Based on Forward Cloud Charge Mechanism," *Int. Rev. Electr. Eng. IREE*, vol. 13, p. 59, Feb. 2018.

[10] A. N. Afandi, Y. Sulistyorini, H. Miyauchi, G. Fujita, X. Z. Gao, and M. El-Shimy, "The Penetration of Pollutant Productions on Dynamic Generated Power Operations Optimized Using a Novel Evolutionary Algorithm," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 7, no. 5, pp. 1825–1831, Oct. 2017.

[11] A. N. Afandi, "Solving Combined Economic and Emission Dispatch Using Harvest Season Artificial Bee Colony Algorithm Considering Food Source Placements and Modified Rates," *Int. J. Electr. Eng. Inform.*, vol. Vol. 6, p. 267, Jul. 2014.

[12] A. Gupta, K. K Swarnkar, and K. Wadhwani, *Combined Economic Emission Dispatch Problem using Particle Swarm Optimization*, vol. 49. 2012.

[13] M. S. Syed, A. R. Syed, and Y. S. Rao, "A novel seeker optimization approach for solving combined economic and emission dispatch," in *2013 International Conference on Power, Energy and Control (ICPEC)*, 2013, pp. 371–376.

[14] A. N. Afandi, I. Fadlika, and Y. Sulistyorini, "Solution of dynamic economic dispatch considered dynamic penalty factor," in *2016 3rd Conference on Power Engineering and Renewable Energy (ICPERE)*, 2016, pp. 241–246.

[15] F. H. Aghdam and M. T. Hagh, "Security Constrained Unit Commitment (SCUC) formulation and its solving with Modified Imperialist Competitive Algorithm (MICA)," *J. King Saud Univ. - Eng. Sci.*, Aug. 2017.

[16] S. Fadıl and B. Urazel, "A security constrained environmental/economic power dispatch technique using F-MSG algorithm for a power system area including limited energy supply thermal units," *Int. J. Electr. Power Energy Syst.*, vol. 56, no. Supplement C, pp. 185–197, Mar. 2014.

[17] L. Wu, J. Gao, Y. Wang, and R. G. Harley, "A survey of contingency analysis regarding steady state security of a power system," in *2017 North American Power Symposium (NAPS)*, 2017, pp. 1–6.

[18] A. N. Afandi, "Weighting Factor Scenarios for Assessing the Financial Balance of Pollutant Productions and Fuel Consumptions on the Power System Operation," *Wseas Trans. Bus. Econ.*, vol. 14, 2017.

[19] A. N. Afandi, "Thunderstorm Algorithm for Assessing Thermal Power Plants of the Integrated Power System Operation with an Environmental Requirement," *Int. J. Eng. Technol.*, vol. 8, pp. 1102–1111, Apr. 2016.

[20] A. N. Afandi and H. Miyauchi, "Improved artificial bee colony algorithm considering harvest season for computing economic dispatch on power system," *IEEJ Trans. Electr. Electron. Eng.*, vol. 9, no. 3, pp. 251–257, May 2014.

[21] K. Chandram, N. Subrahmanyam, and M. Sydulu, "Equal embedded algorithm for economic load dispatch problem with transmission losses," *Int. J. Electr. Power Energy Syst.*, vol. 33, no. 3, pp. 500–507, Mar. 2011.

[22] C. C. Columbus and S. P. Simon, "A parallel ABC for security constrained economic dispatch using shared memory model," in *Controls and Computation 2012 International Conference on Power, Signals*, 2012, pp. 1–6.

[23] B. Y. Qu, Y. S. Zhu, Y. C. Jiao, M. Y. Wu, P. N. Suganthan, and J. J. Liang, "A survey on multi-objective evolutionary algorithms for the solution of the environmental/economic dispatch problems," *Swarm Evol. Comput.*, Jun. 2017.

[24] M. Zheng, X. Wang, C. J. Meinrenken, and Y. Ding, "Economic and environmental benefits of coordinating dispatch among distributed electricity storage," *Appl. Energy*, vol. 210, no. Supplement C, pp. 842–855, Jan. 2018.

[25] L. Lei and W. Shaoqiang, "An Improved Ant Colony Optimization Algorithm Using Local Pheromone and Global Pheromone Updating Rule," in *2016 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS)*, 2016, pp. 63–67.

[26] A. B. Kristoffersen, D. Jimenez, H. Viljugrein, R. Grøntvedt, A. Stien, and P. A. Jansen, "Large scale modelling of salmon lice (Lepeophtheirus salmonis) infection pressure based on lice monitoring data from Norwegian salmonid farms," *Epidemics*, vol. 9, pp. 31–39, Dec. 2014.

[27] A. B. Kristoffersen, L. Qviller, K. O. Helgesen, K. W. Vollset, H. Viljugrein, and P. A. Jansen, "Quantitative risk assessment of salmon louse-induced mortality of seaward-migrating post-smolt Atlantic salmon," *Epidemics*, vol. 23, pp. 19–33, Jun. 2018.

[28] X.-M. Shao, E. H. Lay, and A. R. Jacobson, "Reduction of electron density in the night-time lower ionosphere in response to a thunderstorm," *Nat. Geosci.*, vol. 6, no. 1, pp. 29–33, Jan. 2013.

[29] K. W. Vollset, B. T. Barlaup, H. Skoglund, E. S. Normann, and O. T. Skilbrei, "Salmon lice increase the age of returning Atlantic salmon," *Biol. Lett.*, vol. 10, no. 1, pp. 20130896–20130896, Jan. 2014.

[30] R. L. Holle, "The visual identification of lightning-producing thunderstorm clouds," in *2014 International Conference on Lightning Protection (ICLP)*, 2014, pp. 58–64.

[31] G. Sátori *et al.*, "An Overview of Thunderstorm-Related Research on the Atmospheric Electric Field, Schumann Resonances, Sprites, and the Ionosphere at Sopron, Hungary," *Surv. Geophys.*, vol. 34, no. 3, pp. 255–292, May 2013.

[32] A. N. Afandi, Y. Sulistyorini, G. Fujita, N. P. Khai, and N. Tutkun, "Renewable energy inclusion on economic power optimization using thunderstorm algorithm," in *2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2017, pp. 1–6.

[33] S. L. Haeseker, R. M. Peterman, Z. Su, and C. C. Wood, "Retrospective Evaluation of Preseason Forecasting Models for Sockeye and Chum Salmon," *North Am. J. Fish. Manag.*, vol. 28, no. 1, pp. 12–29.

[34] T. Sabo, "A Survey on Environmental Economic Load Dispatch using Lagrange Multiplier Method," 2012.

[35] A. N. Afandi, "Optimal scheduling power generations using HSABC algorithm considered a new penalty factor approach," in *The 2nd IEEE Conference on Power Engineering and Renewable Energy (ICPERE) 2014*, 2014, pp. 13–18.

[36] L. A. C. Roque, D. B. M. M. Fontes, and F. A. C. C. Fontes, "A multi-objective unit commitment problem combining economic and environmental criteria in a metaheuristic approach," *Energy Procedia*, vol. 136, no. Supplement C, pp. 362–368, Oct. 2017.

[37] M. EL-Shimy, N. Mostafa, A. N. Afandi, A. M. Sharaf, and M. A. Attia, "Impact of load models on the static and dynamic performances of grid-connected wind power plants: A comparative analysis," *Math. Comput. Simul.*, Feb. 2018.

Performance Evaluation of Speed Controller Permanent DC Motor in Electric bike Using Fuzzy Logic Control System

Renny Rakhmawati
Department of Electrical Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
renny@pens.ac.id

Irianto
Department of Electrical Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
irianto@pens.ac.id

Farid Dwi Murdianto
Department of Electrical Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
farid@pens.ac.id, fariddm@ieee.org

Gamal Tabrani Ilman Syah
Department of Electrical Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
gamal.eepis@gmail.com@gmail.com

***Abstract* - The drive system is one of the most important parts on an electric bicycle. One of the drives used is a permanent magnet DC motor. Unfortunately Electric motors pose new problems when applied to electric bicycle mechanics. The problem is divided into 2 parts, the first is the heavy load conditions make the setting point missed from the expected. second is the speed response of the bike system to reach steady state is too slow. Therefore a paper with the title Design and Implementation of Permanent DC Motor for Electric bike Using Fuzzy Method becomes a solution to solve the problems that arise when integrated. Fuzzy method becomes solution to make a control more subtle that can adjust to the load changes. And quickly reach the set point.**

Keywords—Electric bike, Permanent DC motor, Fuzzy

I. INTRODUCTION

The electric motor in this era is very important. The electric motor can do anything to help human activity is like mixer, water pump, electric vehicle and many others [1] They are many kind of electric motor is like separated dc motors, induction motors and dc permanent motors.

DC electric motors have long been applied to electric vehicles. Permanent DC has the advantage of applying an electric bike, which has high torque and has the advantages of not easy to slip and simple to change a speed [2].

Each electric motor has a driver system to drive. The drive system that can be used on a permanent dc motor drive type is a buck converter [3]. One technique used is the technique of setting duty cycle. [4-6] By adjusting the large duty cycle, the electric motor can rotate along with the magnitude of the duty cycle. Many research explain about duty cycle control using Artificial Inteligent [7-12] in DC to DC converter. Electric motors on electric bike tends to be difficult to control when integrated with mechanics. Some problems arise when integrated is like the condition of mechanical loads of basic concern when designing electric bicycles. And the respon speed of time is too slow. From the problems above, the purpose of this paper is to solve the above problems using fuzzy method. In order for electric bikes can go with more stability.

II. MODELING OF ELECTRIC BIKE SYSTEM

In the second chapter are repesent design of electric bike designed system which are block diagram and flowchart.

A. Block Diagram of Electric Bike

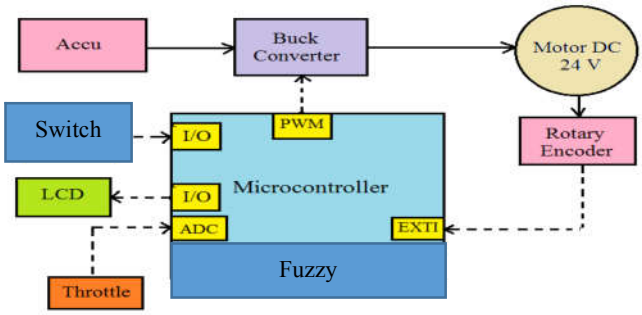


Fig. 1. Block Diagram of electric bike system

Figure 1 present a block diagram of electric bike system with accu as energy source. The energy from the batteries is set from 0-42 volts using the regulation of buck converter (step down DC-DC converter) as a supply of DC motor (DC motor driver). The electric bicycle system uses two modes, manual and automatic. When set manually the system using trothtle (ADC) as the trigger of big duty cycle change in buck converter. The duty cycle can set from 0-100%, it's mean the buck converter can drive speed of DC motor from 0 km/h - maximum rotation km/h. when set automatic mode, the rotary sensor give a data as feed back to microcontroller. The microntroller will give respon and the mode of electric bike system into Fuzzy control

B. *Design of DC Motor*

Design of electric bike is needed to know the DC motor specification. The maximum speed of electric bike is designed 40 km/h. Diameter of wheel is 70 cm, and maximum ability electric bike to withstand load is 120 kg.

$$K = \pi D \quad (1)$$

$$n = \frac{s \times 1000\,m/jam}{K} \quad (2)$$

$$F = Fa + m \times g(\sin\theta^0 + u\cos\theta^0) \quad (3)$$

$$T_L = \frac{F}{2 \times \pi \times \eta} \times \frac{(\pi)(D)}{R} \quad (4)$$

$$Tm = \frac{P}{\omega} \quad (5)$$

Note:

K = Long of circumference
n = rotation
F = Force
Fa = External force
m = massa
g = gravitation
D = diameter
R = Ratio of gear box (1:5)
Tl = Torque of load
Tm= Torque of motor
P = Power of motor

Equation 1 (Long the circumference of a wheel), 2 (Rotation of wheel), 3 (Force) , 4 (Torque load) and 5 (Torqur motor) are required to design a DC motor used. Table I present design value of each calculation.

TABLE I. DATA SPESIFICATION OF ELECTRIK MOTOR DESIGN

K	N	F	TL	TL(new)	P motor
2.198	303.306	149,16	4.73585	6,0795	350 W

C. *Design of Buck Converter (Driver Motor)*

Buck converter is an electronic circuit used to convert dc voltage to dc voltage. The output voltage of buck converter is smaller than the input. Design of electric bike is needed to know the DC motor specification. The maksimum speed of electric bike is designed 40 km/h. The diameter of wheel is 70 cm and maksimum load of electric bike is 120 kg (mechannic and human).

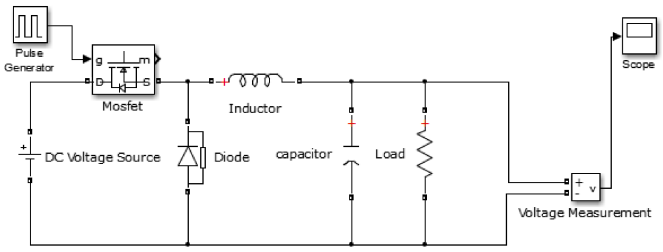


Fig 2. Buck converter circuit

$$V_{obuck} = V_s \max \times D \quad (6)$$

$$L = \frac{1}{f} * (Vs\max - Vo) * \left(\frac{Vo+Vf}{Vs\max + Vf}\right) * \left(\frac{1}{\Delta iL}\right) \quad (7)$$

$$C = \frac{1-D}{8 \times L \times (\Delta V/Vo) \times f^2} \quad (8)$$

Note:
Vin : Input Voltage
Vo : Output Voltage
L : Inductor
C : Capacitor

Equation 6 (Output voltage of buck converter design), 7 (Inductor design) and (capacitor design) are required to design a DC motor used. In this paper buck converter used to control input voltage of DC motor (driver of DC motor). Table II is present design value of buck converter and figur 3 is a hard ware of buck converter.

TABLE II. DESIGN VALUE OF BUCK CONVERTER

Vin	Vo	L	C
42	24	66.577	100,58 uF

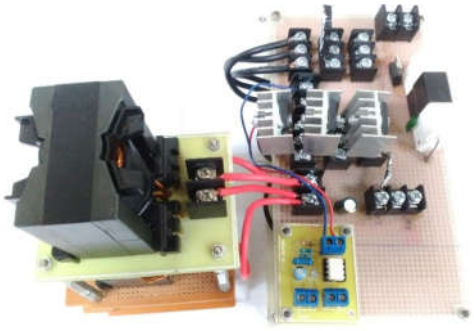


Fig. 3 . Hardware of buck converter

D. Electric Bike Design

Electric bike is needed mechanic design. It is necessary to design the bike to be integrated with electrical components and can be easily driven. Mechanical design on electric bicycles is generally as same as with other bikes that exist in the market, but this electric bike using a resuffle design of bike. Figure 4 is the electric bike mechanic design on electric bicycles is generally the same as other electric bikes that exist in the market, but this electric bike using a design reshuffle of wind vehicles with 350 Watt Permanent DC motor.

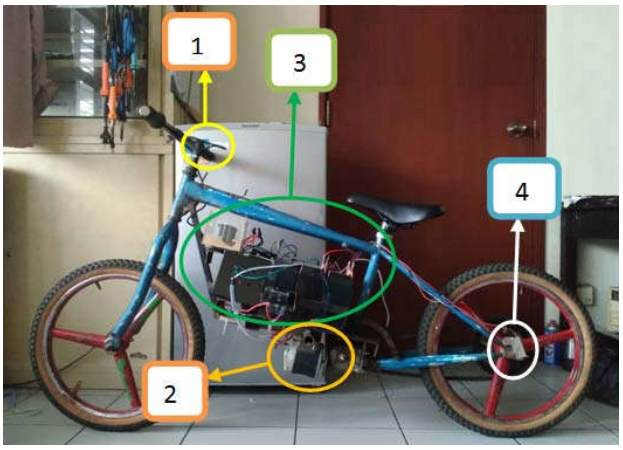


Fig.4. Electric bike design

Note:

1. Throttle
2. Permanent DC motor
3. Electric component
4. Rotary encoder sensor

er

E. Design of Fuzzy Logic Controller

Fuzzy logic design of this paper is use to set value of dutycycle. Figure 5 is the represent of diagram blog of speed control using fuzzy logic.

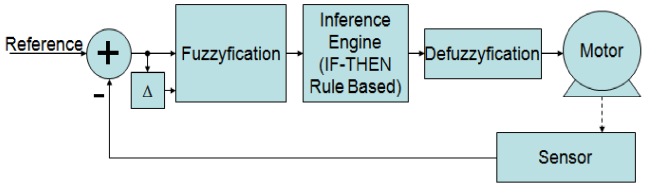


Fig. 5. Diagram blog of fuzzy contol design

There are two input data, error and delta error

1. Variable of Input Error (E)

Input error numerik is like figure 6, has range from -80 to 80

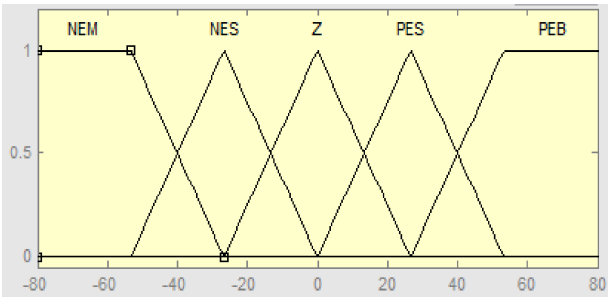


Fig.6. Error of membership function 5x5

2. Variabel Input of Delta Error (dE)

To get a smooth tracking, the input change in error is like figure 7 the delta error set same as membership error function from -80 to 80.

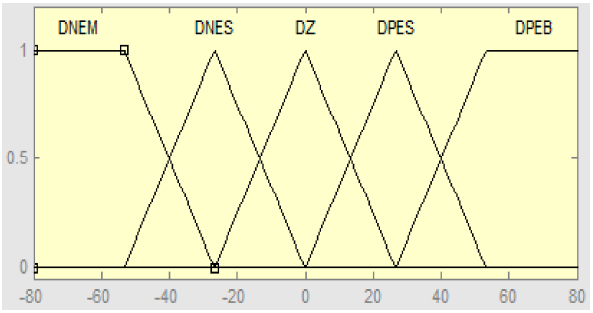


Fig.7. Error of membership function 5x5

3. Rule base

Rule base is linguistic rule of respon. Rule base is important to get accordingto which desired. Table III is represnt of rule base

TABLE. III. RULE BASE OF FUZZY LOGIC CONTROLLER

	NEM	NES	Z	PES	PEB
DNEM	DL	DL	DM	DS	NC
DNES	DL	DM	DS	NC	IS
DZ	DM	DS	NC	IS	IM
DPES	DS	NC	IS	IM	IL
DPEB	NC	IS	IM	IL	IL

Note:
DL = Decrease Large
DM = Decrease Medium
DS = Decrease Small
NC = Not Change
IS = Increase Small
IM = Increase Medium
IL = Incrase Large

4. Output Variable (Singletone)
Fuzzy logic output is singletone. Figure 8 is singletone of fuzzy logic

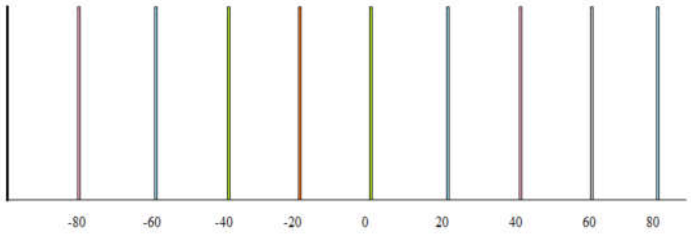


Fig. 8. Output singletone

III. Simulation and Experimental Results

In the third chapter are comparing simulations on matlab simulink with real time results on measurements on the output speed respon.

A. *Electric Bike whit Speed Control Simulation*

Simulation of speed control was done using Fuzzy method by matlab simulink. With this simulation is expected that output speed output of electric bike is constant. Figure 9 present transfer function block of matlab simulink for electric bike system. Figure 10 is a chart of output speed respon of electric bike without control, figure 11 is a chart of output respon of electric bike whit load disturbance (without control) and figure 12 is a chart of output respon of electric bike whit load disturbance (Fuzzy controlled).

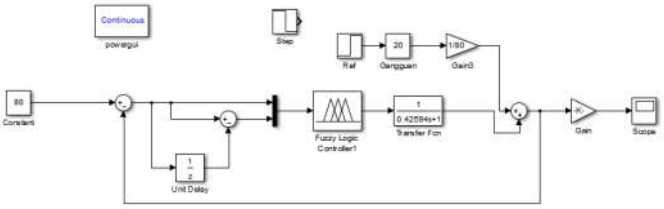


Fig. 9. Simulasi boost converter with constant voltage control

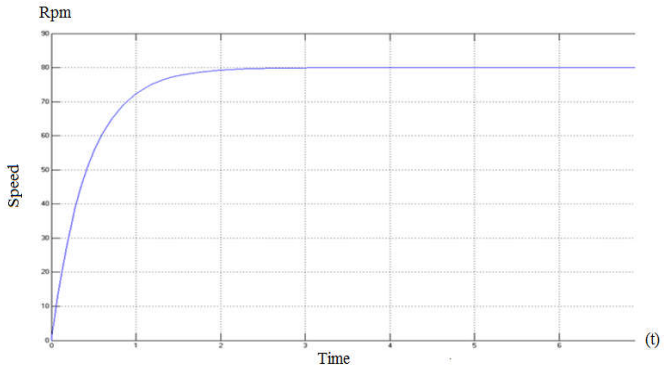


Fig. 10. Speed respon in openloop mode

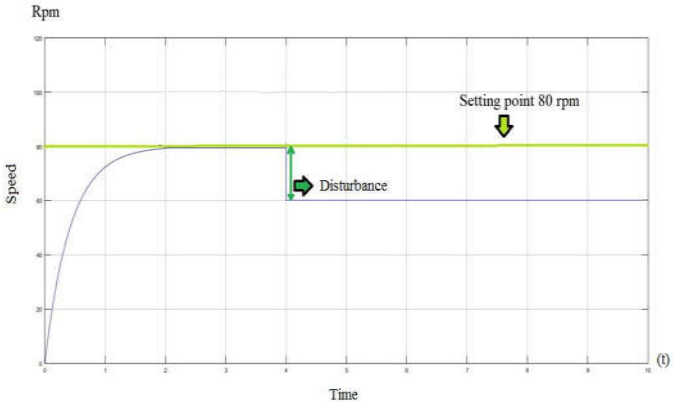


Fig. 11. Speed respon in openloop mode whit disturbance

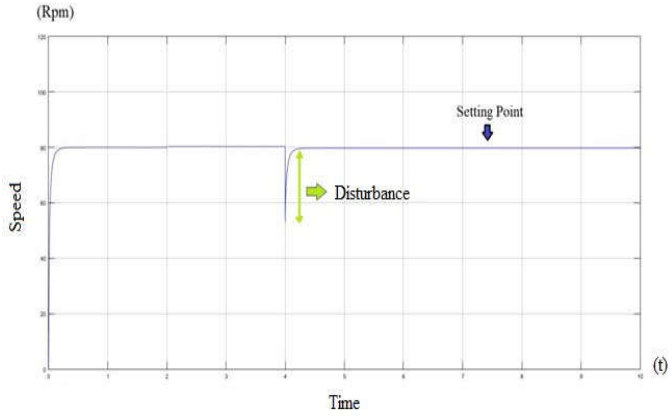


Fig.12. Speed respon in closeloop (FUZZY control) mode whit disturbance

B. *Elctric bike system with FUZZY Control Experimental*

The control speed of electric bike using FUZZY logic control aims to enhance a time respon to reach setting point. Comparison of open loop and close loop (Fuzzy control) responses are presented in tables 4 and 5. In this experience the information of load is human = 50 Kg, mechanic = 68 Kg.

TABLE IV. ELECTRIC BIKE TEST WITHOUT CONTROL

Rotation of wheel (Rpm)	Time (s)
0	0
27,13	1
26,1	2
28,16	3
54,31	4
64,37	5
70,40	6
75,40	7
75,43	8
83,48	9
83,48	10

TABLE V. ELECTRIC BIKE TEST USING CONTROL (FUZZY)

Rotation of wheel (Rpm)	Time (s)
0	0
32,18	1
53,30	2
69,40	3
81,46	4
80,46	5
79,45	6
78,45	7
81,46	4
80,46	5
79,45	6
78,45	7

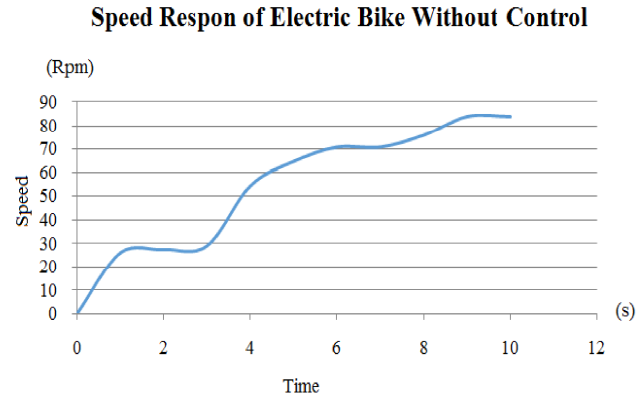


Fig. 13. Speed respon without control

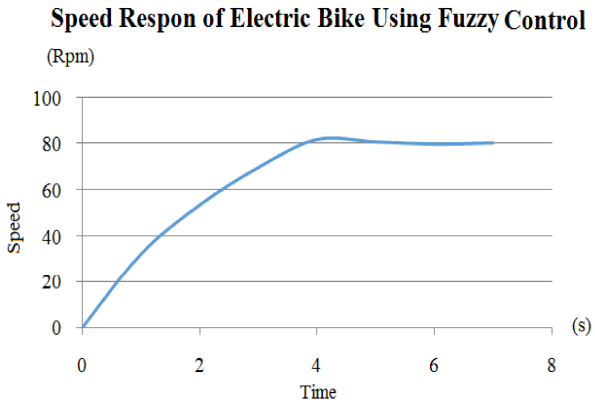


Fig. 14. Speed respon with fuzzy control

After get a speed respon, the next experiment is give a disturbance when the system work. Table 6 and table 7 is the realtime data of disturbance respon. The load condition of this experiment is human = 50 Kg, electric bike mechanic = 68 kg and the disturbance load = 7 Kg.

TABLE VI. ELECTRIC BIKE TEST WITHOUT CONTROL

Rotation of wheel (Rpm)	Time (s)
0	0
27,13	1
26,1	2
28,16	3
54,31	4
64,37	5
70,40	6
75,40	7
75,43	8
83,48	9
83,48	10
83,48	11
83,48	12
70.51	13
70.51	14
70	15
70.23	16
70.23	17

Based on the test data, the graph of the response by using fuzzy logic control and without fuzzy logic control can be seen on figure 13 and Figure 14.

TABLE VII. ELECTRIC BIKE TEST USING CONTROL (FUZZY)

Rotation of wheel (Rpm)	Time (s)
0	0
32,18	1
53,30	2
69,40	3
81,46	4
80,46	5
79,45	6
78,45	7
70,51	8
70,53	9
75,45	10
77,45	11
77,45	12
77,45	13
77,45	14

Note:
Green color has mean that the respon is gave a disturbance.

Figure 15 is open loop respon without disturbance and figure 16 is the represent of close loop speed respon with disturbance.

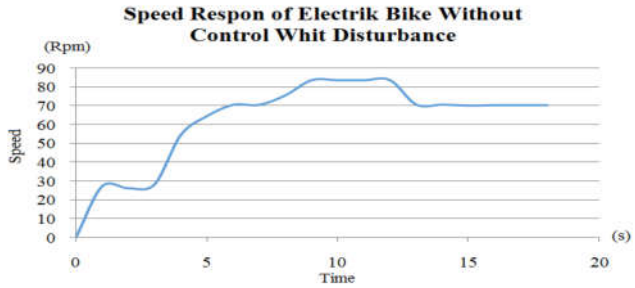


Fig. 15. Open loop speed respon with disturbance

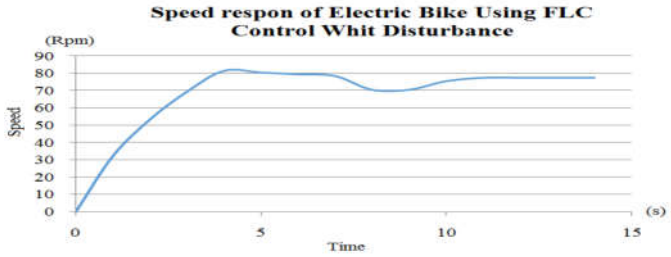


Fig. 16. Close loop speed respon with disturbance

IV. CONCLUSION

Simulation result show that the fuzzy control can be applied in electric bike system with the time respon increase 5 second faster than uncontrol system. The fuzzy can keep the set point although the system gave a disturbance with increase load.

References

[1] Jaya A., Purwanto E., Fauziah M.B, Murdianto F.D, Prabowo G, Rusli M.R, "Design of PID-Fuzzy for Speed Control of Brushless DC Motor in Dynamic Electric Vehicle to Improve Steady-State Performance" (IES-ETA) International Electronics Symposium on Engineering Technology and Applications, Surabaya, 2017.

[2] Jaya A, Wahjono, Rusli M.R, Purwanto E, Murdianto F.D, Fauziah M.B, "Design and Simulation of Sensorless BLDC Motor Drive Using Flux Linkage Increment Based on the line-to-line BEMF for Electric Vehicles" (IES-ETA) International Electronics Symposium on Engineering Technology and Applications, Surabaya, 2017.

[3] D. W. Hart, "The Buck Converter," dalam Power Electronics, New York, McGraw-Hill, 2011, pp. 198.

[4] M. Akil, I. Nurtato "A DC Motor Speed Control Using The LPC-ANFIS Speech Recognition System " *Indonesia, Hasanuddin Univercity, IEEE, 2017*

[5] H. Chaundary, S. Khatoon and R. Singh "ANFIS Based Speed Control of DC Motor"New delhi, India, IEEE, 2016

[6] J. Lee , J. Jiang and Y. Sun " Design and Simulation of Control Systems for Electric-Assist Bikes" National Taipei University of Technology (Taipei Tech.), Taipei, Taiwan,IEEE 2016

[7] Murdianto F.D, P. Ontoseno, P. Ardyono, "Modeling and simulation of mppt-bidirectional using adaptive neuro fuzzy inference system (ANFIS) in distributed energy generation system," IEEE International Seminar on Intelligent Technology and Its Applications (ISITIA), 2015.

[8] Soedibyo, Murdianto F.D, Suyanto, A. Mochamad, P. Ontoseno,"Modeling and simulation of mppt SEPIC combined bidirectional control invers KY converter using ANFIS in microgrid system," Indonesian Journal of Electrical Engineering and Computer Science, vol. 1, pp. 264-272, February 2016.

[9] Suyanto, Murdianto F.D, P. Ontoseno, S. Adi, S. Rony,"Wind-PV hybrid system modeling using bidirectional converter with MPPT-dual adaptive neuro fuzzy inference system (ANFIS) in microgrid isolated system," Journal of Engineering and Applied Sciences, vol. 11, pp. 2353-2359, December 2016.

[10] Rakhmawati R, Wahjono E, Suhariningsih, Murdianto F.D, Andriawan D,"Design and Implementation Boost Converter with Constant Voltage in Dynamic Load," International Seminar on Application for Technology of Information and Communication (iSemantic), 2017

[11] Efendi M.Z, Murdianto F.D, Setiawan R.E,"Modeling and simulation of MPPT sepie converter using modified PSO to overcome partial shading impact on DC microgrid system," International Electronics Symposium on Engineering Technology and Applications (IES-ETA), 2017.

[12] Murdianto F.D, Nansur A.R, Hermawan A.S.L,"Modeling and simulation of MPPT coupled inductor sepie converter using Flower Pollination Algorithm (FPA) Method in DC mierogrid system," International Electronics Symposium on Engineering Technology and Applications (IES-ETA), 2017.

[13] F.D. Murdianto, M. Z. Efendi, R. E. Setiawan and A. S.. L. Hermawan, "Comparison method of MPSO, FPA, and GWO algorithm in MPPT SEPIC converter under dynamic partial shading condition", International Conference on Advanced Mechatronics, Intelligent Manufacture, and Industrial Automation (ICAMIMIA), 2017

[14] F.D. Murdianto, A. S.. L. Hermawan, A. R. Nansur and R. E. Setiawan, "Comparison method of flower pollination algorithm, modified particle swarm optimization and perturb & observe in MPPT coupled inductor sepic converter on DC microgrid isolated system", International Conference on Advanced Mechatronics, Intelligent Manufacture, and Industrial Automation (ICAMIMIA), 2017

Internet of Things on monitoring and control system in server area

Ahmad Roihan, Ferry Sudarto, Trengginas Cahyo Putro
Computer System Department
STMIK Raharja
Tangerang, Indonesia
ahmad.roihan@raharja.info, ferry.sudarto@raharja.info, trengginascahyoputro@raharja.info

***Abstract*—The high demand for mobile information to achieve the effectiveness of a job is the reason for improving technology at this time. The existence of a mobile monitoring system is needed to replace the monitoring activities that are still done manually. Monitoring server area managed by the company itself is still done manually which is considered less effective and the lack of automatic security in the event of short circuit and fire that can lead to losses from damage to the server device and the cessation of the company's operational activities. This research aims to create a tool that is able to provide information about the temperature and condition of server area in mobile and real time. This research also used data collection method by observation that ended with test method. The monitoring system uses the concept of Internet of Things (IoT) and sending message with SMS gateway. The device input comes from the MQ-02 Sensor, DHT22 Sensor and Wemos D1 with the resulting output of LED, Buzzer, selenoid valve, sim900 module and web display of ubidots. Then the performance of this tool is to detect the temperature and levels of smoke in the server area that is sent to the web ubidots, if the smoke detected fire extinguisher faucet will automatically open with LED and buzzer alarm and give message to the user's mobile phone number. So that with the monitoring tool is able to facilitate the work of officers when performing monitoring activities on server area and provide a sense of security to the company.**

Keywords—Monitoring; Server area; Internet of Things; SMS Gateway.

I. INTRODUCTION

After more than 20 years of development, Internet of things has a lot of applications in the actual scene, which greatly facilitates people's work and life [1]. The role of computerized system using server services is very important in the development of business in the era of globalization as now, the server as a data center is considered capable of improving the effectiveness and efficiency of a job because with the integration of data on a server through the network can replace the processing or sending data manually.

Along with the growing awareness of information technology some companies choose to integrate its operational activities using a computerized system connected to the server device. The number of purposes on the server device requires the company has its own server area in order to be protected from unintentional errors such as human error and deliberate activity by irresponsible people.

The existence of monitoring activities on the server area is very important to maintain the stability of the server device inside it by monitoring the temperature and ensuring the absence of electrical short circuit or fire in the server room. However, monitoring activities performed manually are considered less effective because it takes a lot of time and energy.

The high demand for mobile information to achieve the effectiveness of a job, some companies require a monitoring tool that can replace the monitoring activities performed manually. Therefore, Integrasi Solusi Prima corporation which is a company engaged in the field of information technology since 2005. Operationally carried out by the network engineer is monitoring the network and sever area. In the monitoring of server area managed by the company, the reliability in monitoring and security is less optimal because it is still done manually and recorded periodically by looking at the conditions in the server room, the activity is still considered less effective because of the lack of engineers who vigil with various jobdesk done, the length of the permission stage to enter the server room and the absence of real time temperature monitoring system.

Lack of security in the server room in the event of a short circuit and fire into one of the company's concerns because it can lead to significant losses from damage to the server device and the cessation of the company's operational activities.

This research aims to make monitoring and control system in real-time and mobile on server area, then automatic fire safety that can facilitate the work of officer while doing monitoring activity on server area and give security to company side.

In this research, there are problems that can be formulated among others that is needed to design the temperature and smoke detection circuit with wemos in server room at company, then tool designed in server room can be monitored in real time and then make fire safety automatically on the server room in case of short circuit and fire.

The scope of this study is limited to monitoring process using DHT22 sensor to detect temperature and MQ-02 sensor to detect smoke. Then the tool concept of Internet of Things (IoT) with SMS Gateway as a means to deliver real-time temperature results on web ubidots and notification in the form of messages on mobile phones. Next safeguard against fire early using selenoid valve as automatic fire extinguisher faucet when detected smoke and fire.

II. CONCEPTS

Research method is a way or steps used in doing a research. A study has a specific research design. The design describes the procedures or steps to be taken, the timing of the study, the data source and the condition of the data collected and in what way the data is collected and processed for use in report generation.

There are several research methods used in making this research, the first method is the method of data collection. This research through the early stages of making observations to collect data that can be supporting research information in analyzing and problem solving. The next stage is to conduct a survey conducted on the correspondent or the parties involved in the research. Correspondent is given a question related to the conditions and problems in the monitoring of the server area. Another method of data collection is the literature review of various materials either through the literature of books and writings or scientific articles from several international journals as a complement to research data.

After getting various data, the next method is the method of analysis. It can be implemented to create tools that suit the needs of stakeholders. Based on the results of the analysis that has been done on the running system, there are some problems encountered by the server engineer in the monitoring of server space such as monitoring the temperature is still done manually, more time and energy, temperature check can not be done precisely on time, not equipped with a notification if the temperature is above the standard and detected smoke, there is no fire protection automatically.

In the design method, there are several requirements specifications needed to assemble a monitoring and control system based on IoT and SMS gateways. The main components are DHT22 sensor and MQ-02 sensor which both as input, wemos as data processor from input signal to determine output part task, then SIM900 module, LED, buzzer, selenoid valve, mobile to receive SMS (Short Message Service) in which all of these components serve as outputs. After the tool is finished assembled the next step is to do the encoding of wemos development board.

Research analysis in the selection of DHT22 sensors is appropriate in measuring the temperature and humidity of the surrounding air [1]. The selection of MQ-02 usage that can detect smoke and airborne gas concentrations in the air and then accompanied by the use of SIM900 due to its very precise functioning according to previous research experience [2]. Buzzer and LED placement was determined according to the needs based on observation. Then selenoid valve is chosen because this tool is a valve driven by electrical energy, this tool has a coil as a driver that serves to move the piston that can be driven by AC and DC voltage so that this tool can be used as an automatic valve for fire extinguisher exit.After the tool has been able to run the last method is a test method to ensure the system has been going well, then done the test scenario to determine the overall performance of the tool which should be appropriate to what has been planned.

This research is a development of some previous research related [3]. The research object is limited only in server area in Integrasi Solusi Prima corporation. The current research is related to the design of monitoring and control system in server area. Several research have been conducted related to this research among which is done by Xiaodong Wang which firstly to formulate the problems of sensor placement for hot server detection in a data center as constrained optimization problems in two different scenarios and then propose a novel placement scheme based on computational fluid dynamics (CFD) to take various factors, such as cooling systems and server layout, as inputs to analyze the thermal conditions of the data center. Based on the CFD analysis in various server overheating scenarios, this research apply data fusion and advanced optimization techniques to find a near-optimal sensor placement solution, such that the probability of detecting hot servers is significantly improved [4].

Other research from D. Kalyanraj which proposes a low cost Wind turbine monitoring and control system with data logging facility. By the use of this proposed system, Parameters of the particular wind plant like amount of Power generation, Magnitude of instantaneous voltage and currents, level of vibration, Turbine speed, Humidity and Temperature can be monitored and used anywhere. In the proposed system Turbine control is adopted based on the turbine vibration level and IoT is implemented by using Raspberry Pi and arduino microcontrollers [5]. Next research from Wanda Vernandhes who has been discussing about Agricultural technology design with aquaponics is also using the concept of Internet of Things because the information from the sensor and control actuator values can be accessed through applications installed on the smartphone from anywhere with the Internet connection. Agricultural cultivation technology with indoor aquaponic agricultural technology provides an alternative for anyone who has no land for farming but can still conduct business activities that can be used as an adequate source of food or income [6].

The design of the system proposed in this research is to create a monitoring and control system in the server room using IoT and SMS gateway that can be accessed by mobile, then that is expected to help and provide convenience to the engineer in monitoring the server area to provide a sense of security for the company and the owner of the server device.

III. SYSTEM ARCHITECTURE

Monitoring activities undertaken for the progress of an ongoing project are aimed at knowing the process in the activity. In the design stage of this tool required components will be formed that resemble the use of monitoring system and control the server space, the design of the tool in this stage includes several components, the SIM900 which serves as the receiver and notification message sender, DHT22 sensor that serves to detect the temperature, MQ-02 sensor that serves to

detect smoke, wemos D1 which serves as a controller of the circuit which is useful for communication to ubidots server via wifi network that will be displayed on web ubidots, then LED and buzzer for notification directly to the officers who are in the location when electrical short circuit, burning device and temperature above standard or detected smoke then selenoid valve will open the fire extinguisher faucet.

This study deals with IoT, issues faced with the IoT are realized that most of the challenges are privacy or data confidentiality and integrity. Later in the study have found a way of securing information shared between two or more interactive IoT devices. The main objective of this project was to develop a strategy of securing the world of IoT technology [7].

Figure 1 illustrates the input process that occurs when the MQ-02 sensor detects smoke, the DHT22 sensor detects the temperature, SIM900 receives the SMS (Short Message Service) and sends a signal to the development board wemos to then be detected from the sensor.

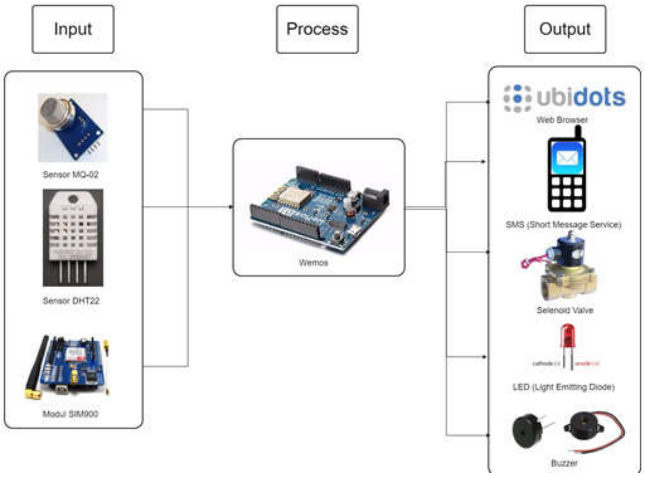


Fig. 1. Ubidots server side

In the process section, the wemos will process the information received from the sensor to be processed into a data. Wemos will determine the work to be assigned to the output section.

After the input and process have been done on the tool then the output can be generated, then the result of the output section consisting of the temperature section will be displayed on web ubidots that serves for detection and control on the server room. In addition to displaying on web ubidots, temperature and smoke detection can be sent via SMS (Short Message Service). Other output sections include LEDs and buzzers as direct notifications on the server room when temperature is detected above standard, and also equipped by a selenoid valve as an automatic fire extinguisher faucet when it detects smoke in the server room.

Stages of hardware design required some electronic components as a support system can run in accordance with its function. The following components of electronics required in performing hardware design such as laptops used as a tool to write the program into the wemos deveopement board and to design the electronic component circuit scheme which is required in making the tool. the mobile phone acts as the receiver of the message notification from the SIM900 module gateway SMS. Wemos D1 is a development board that features wifi as a support for wireless data transmission into a network. SIM900 module served as SMS gateway that will send notification message on handphone officer. DHT22 sensor that served as temperature detector. The MQ-02 sensor serves as a smoke detector. Selenoid valve that serves as a controllable automatic faucet. Solder as a tool to melt lead to connect several components. Lead is a very good conductor serves as a tool to connect components in order to form a series of expected tools. Other components are LED, Buzzer, Adapter, PCB, Resistor and Jumper.

The design of the hardware required a picture that contains the needs of what is used to create monitoring tools and control the server room with their respective functions, in order to easily understand the researcher makes the block visual diagram in Figure 2 below,

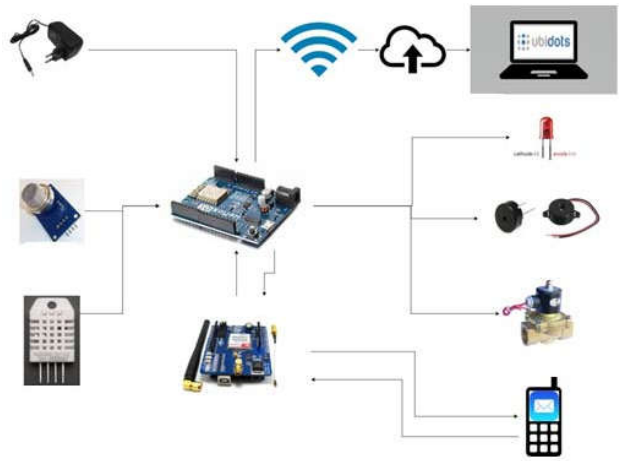


Fig. 2. Block visual diagram architecture

The adapter will supply the electric current voltage as a power source for the wemos and the module SIM 900. Wemos will work if there is an incoming signal from the sensor to be processed into data then determine which devices will work and receive data. MQ-02 serves to detect smoke levels in the server room in the event of fire caused by electrical short circuit and other smoke triggers. DHT22 serves to detect the temperature on the server room space. Wemos communication process in sending data via internet that is using wifi connection to connect with web ubidots. Further data will be sent via the internet. Web ubidots will receive data from wemos D1 and display it on the user's web browser. The LED indicator will turn on as a direct visual notification on the server room when the temperature is above standard. Buzzer indicator will sound as direct notification on server space when temperature above standard and smoke detected. Then selenoid valve serves as an automatic fire extinguisher faucet when it detects smoke in the server room. Furthermore, the SIM 900 Module acts as the receiver of instruction messages and sends a notification message from the monitoring tool and control the

server room to the mobile phone number which is already registered in the system.

The next stage is designing software that requires some supporting applications as a support system can run in accordance with its function. As for some supporting application in system design that is Arduino IDE (Integrated Development Enviroment) software, ubidots website and web browser.

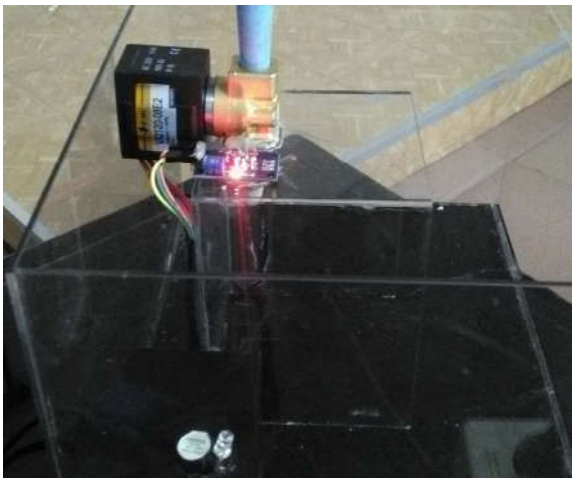


Fig. 3. Architectural model

In designing this system architecture will be done by making a model for the implementation of it can run properly and in accordance with the needs of the company based on existing problems. Therefore, in addition to building architectural model as in Figure 3 also made scenarios of testing against some sensors so that no errors in the calculation of server space. Installation of any tool will be adjusted to be limited only to the server rack in the server room. As if there is electrical short circuit outside the rack area, then this can be prevented in another way and developed in other research.

IV. TESTING

The testing process performed by the DHT22 sensor to detect the temperature in a room is done by connecting the data pins of the DHT22 sensor connected to the D2 pin on Wemos D1, then programming to obtain temperature data from the DHT22 sensor. The working principle on the DHT22 sensor circuit is the sensor will read the temperature in the server room and wemos will send data from the sensor to the web ubidots.

Further testing process on the sensor MQ-02 which aims to detect smoke in the server room in real time by connecting the pin A0 on the sensor MQ-02 to the analog pin 0 in wemos. The working principle of sensor MQ-02 is sensor will detect smoke then wemos will give instruction to selenoid valve to open, sim900 will give message notification and wemos will give information to web ubidots that server room detected smoke.

Tests performed on the SIM900 module to check the network on the SIM900 module is by connecting the digital pins 7 and 8 in the module to the digital pins 7 and 9 on the wemos. The working principle of the SIM900 module is that when the MQ-02 sensor detects smoke in the server room, the SIM900 sends a message notification and gives instructions to several parts so that the selenoid valve can open, the LED and buzzer will be lit.

In selenoid valve testing, the fire extinguisher faucet will open when wemos development board gives instructions. The power circuit on the selenoid valve is connected to NC (normally closed) and then on the relay pin connected to the pin on the wemos.

An authorized person has been informed with web application [8]. It's like this research, ubidots web display testing stage displays the temperature status with the DHT22 sensor and the smoke level with the MQ-02 sensor on the server area sent from the development board wemos D1 to web ubidots automatically and in real time. The color in the smoke graph on the web ubidots gives a description if it is gray then indicates no smoke detection and vice versa if it is green then gives a signal that the smoke detected in the server room. Then in the temperature graph gives the temperature degree information with the Celsius unit. Figure. 4 below is the result of smoke and temperature detection that has a negative result on the server room, while Figure 5 is the result of the detection of smoke and temperature above the standard on the server room. Both are displayed by the dashboard on the web ubidots.



Fig. 4. Dashboards ubidots which was detected negatively



Fig. 5. Dashboards ubidots which was detected above the standard

Based on testing as a whole there are two testing methods that is software by using Arduino IDE application and hardware using development board wemos D1. After testing with Arduino IDE software and wemos D1, it is necessary to test again so that wemos can provide sensor data on web ubidots and test SIM900 module in order to send notification message to user's mobile phone number with predefined message format so that from research in software and hardware gets good results.

V. CONCLUSION

Based on the results of system design that has been described previously there are several subjects that can be concluded. On the server room monitoring activities in the

company which has been running is considered less effective because it is still done by manually recording by visiting the server room. Therefore, research needs to be done by creating monitoring and control tool on server area based on Internet of Things (IoT) and SMS gateway in order to assist in monitoring activity of server area.

In designing the monitoring system server area is needed attention to two conditions in the server room that is the condition of detecting the temperature with DHT22 sensor and the detection of smoke, short circuit or fire with the sensor MQ-02. Wemos D1 is used as a control on the output part such as buzzer, LED, selenoid valve (automatic fire extinguisher faucet), displaying on web ubidots and SMS gateway as message notification using SIM900 module. The presence of monitoring and control tools on the server room using IoT and SMS gateway is able to simplify the monitoring activities carried out by the network engineer and provide a sense of security for the company because it can be watched in real time and mobile.

ACKNOWLEDGMENT

The authors would like to thank STMIK Raharja and Integrasi Solusi Prima corporation, in particular for the support which has provided an opportunity to conduct research in industrial areas. Then to colleagues who help or provide support related to the research undertaken.

REFERENCES

[1] Z. Tu, C. Hong and H. Feng, "EMACS: Design and implementation of indoor environment monitoring and control system," IEEE Computer and Information Science (ICIS), Wuhan, pp. 305-309, 2017.

[2] A. Roihan, A. Permana and D. Mila, "Monitoring Kebocoran Gas Menggunakan Mikrokontroler Arduino UNO dan ESP8266 Berbasis Internet of Things," ICIT (Innovative Creative and Information Technology), vol. 2, no. 2, pp. 170-183, 2016.

[3] H. Ali, J.K. Soe and S.R. Weller, "A real-time ambient air quality monitoring wireless sensor network for schools in smart cities," International Smart Cities Conference (ISC2), Guadalajara, pp. 1-6, 2015.

[4] X. Wang, X. Wang, G. Xing, J. Chen, C. Lin and Y. Chen, "Intelligent Sensor Placement for Hot Server Detection in Data Centers," IEEE Transactions on Parallel and Distributed Systems, vol. 24, pp. 1577-1588, 2013.

[5] D. Kalyanraj, S.L. Prakash and S. Sabareswar, "Wind turbine monitoring and control systems using Internet of Things," IEEE 21st Century Energy Needs - Materials, Systems and Applications (ICTFCEN), Kharagpur, 2016.

[6] W. Vernandhes, N.S. Salahuddin, A. Kowanda and S. P. Sari, "Smart aquaponic with monitoring and control system based on iot," IEEE Informatics and Computing (ICIC), Jayapura, 2017.

[7] G. Matsemela, S. Rimer, K. Ouahada, R. Ndjiongue, Z. Mngomezulu, "Internet of things data integrity," IEEE IST-Africa Week Conference, 2017.

[8] G. Çetin, M.Ş. Özdemir, S. Karadaş, Y. Yılmaz, "IEEE 802.15.4 / ZigBee based wireless sensor network design for monitoring server rooms," IEEE International Conference on Computer Science and Engineering (UBMK), Antalya, pp. 565-570, 2017.

Soft Starting & Performance Evaluation of PI Speed Controller for Brushless DC Motor Using Three Phase Six Step Inverter

Renny Rakhmawati
Electrical Engineering Department
Electronics Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
renny@pens.ac.id

Farid Dwi Murdianto
Electrical Engineering Department
Electronics Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
farid@pens.ac.id, fariddm@ieee.org

Muhammad Wildan Alim
Electrical Engineering Department
Electronics Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
muhammadwildanalim@yahoo.co.id

Abstract—Along with the developments that occur on the use of DC brushless motor as a driver. Brushless DC motor (BLDC) has good performance, high speed, responsive, reliable and more efficient. Because of the advantages of the BLDC motor is often used in industries, electric vehicle propulsion and so forth. However, the use of BLDC motor has a problem that is the use of a very high starting current starting motor so as to cause energy stored in batteries or other energy storage is very quickly exhausted and reduced life time motor arena transient currents. It needs a control to make the starting BLDC motor to reduce the starting current spike by stepping step duty cycle inverter slowly until it reaches duty setting. In addition to the BLDC motor needs a control to set reach the set point rpm faster. Therefore, this paper describes the use of BLDC starting motor & BLDC motor speed setting with PI control. It will show the difference of starting current performance and BLDC speed when using the method.

Kata Kunci—BLDC motor; Soft Starting; PI Controller.

I. INTRODUCTION

Along with technological developments, the need for a motor that has efficiency, high speed torque and can be varied, as well as low maintenance costs increased. It's just that the motors used in general today, namely DC motors and induction motors have not been able to meet those needs.

DC motor has a high efficiency because the use of DC voltage on the rotor to drive the motor, but the DC motor has a high maintenance cost. This high maintenance cost arises from the use of a brush in DC motor commutation. Brush on DC motor is quickly damaged because at the time of the motor rotates, the brush will arise sparks due to the process of commutation.

Induction motors have higher speed and lower maintenance costs compared to DC motors. This happens because the induction motor does not use the Brush to support commutation. However, induction motor has lower efficiency compared to DC motor and fast heat. This occurs because the rotor made of the coil is induced by a rotating magnetic field of the stator resulting in a current on the rotor. The current on the rotor coil causes the rotor to become hot because there is power dissipation. In addition, because this motor can only work when the rotor coil is switched currents and becomes a magnet, the rotation speed of the rotor will be much lower than the rotational speed of the stator magnetic field. This causes induction motor efficiency to tend to be lower than DC motor
[1].

The construction of a DC brushless motor consists of a stator made of coils and a rotor made of permanent magnets. The main voltage source of the BLDC motor is the DC voltage, BLDC motor has a trapezoidal EMF back. In order for the BLDC motor to work, a magnetic field of stator rotation is required. To obtain a rotating magnetic field the stator required a 3 phase AC voltage source on the motor stator. Therefore a 3-phase inverter is used to convert DC voltage into 3 phase AC voltage. Based on the form of 3 phase AC voltage generated by the inverter there are 2 methods in controlling the BLDC motor inverter that is by using the six-step method and the sinusoidal PWM method [9-16].

The six-step method is a frequently used method of controlling BLDC motors. This happens because this method is easy to implement and has a simple algorithm. The waves generated from this method are square or trapezoidal. To form trapezoidal wave or 3 phase square wave used 3 pieces of six-step algorithm each different one step (60 degree) between one algorithm with other algorithm. BLDC motor speed can be adjusted by using PWM technique as input from inverter and varying DC voltage as input for DC brushless motor. Motor speed setting can be done by changing the amount of input voltage provided on the motor BLDC in accordance with the desired motor speed. Motor speed setting can also be done by changing the pulse duty cycle. The pulse that varies its duty cycle can determine the speed of the BLDC motor. The magnitude of the amplitude and frequency of the pulse is fixed, while the magnitude of the duty cycle can be changed according to the desired speed of the BLDC motor.

In order for the motor to work with torque and constant speed, it is necessary to select the correct timing of commutation change in the BLDC motor control [8] both for the six-step method and the sinusoidal PWM method. In the BLDC motor used encoder or 3 hall sensors to determine the timing of commutation changes in BLDC control. In general, the encoder is used to determine changes in the selection of commutation time. But since the encoder is fixed (irreversible), an encoder may not necessarily be applied to other motors. This happens because if the motor has a different number of poles, the encoder used must also be different. This is different from the hall sensor, is when the number of poles of the motor changes, the location of the hall sensor can be easily changed.

The method used in this paper is to use a six step. This inverter duty cycle improvement is done with step by step until it reaches the desired duty. The duty cycle in this study was obtained from comparing the step 2 signal with the sawtooth signal. Thus, this method can be used to reduce the overshoot of the starting current on the BLDC motor. In addition, speed control in this paper using PI method. With the PI method it is expected that the response output of the BLDC motor will accelerate to reach the specified set point.

I. INVERTER SIX STEP & BRUSHLESS DC MOTOR

A. Inverter Six Step

The inverter is a converter to convert DC to AC, which has the function of converting DC voltage to an AC frequency voltage which can be adjusted. The output variable voltage can be adjusted by adjusting the input voltage. If the input voltage is constant, then we can adjust the ouput voltage by adjusting the gain of the inverter which is usually done with PWM control. This inverter gain is defined as the ratio of the AC output voltage to the DC output voltage.

The inverter division is based on the number of phases, divided into two, ie single phase inverters and three phase inverters. The one phase inverter is a one-phase DC to AC converter, whereas a three-phase inverter is a three-phase DC to AC converter. Based on the input used, the inverter type includes VSI (Voltage Source Inverter) and CSI (Current Source Inverter). In VSI the input source is a voltage while at CSI the input source is a current.

Output signal output wave depends on the commutation signal. While the commutation arrangement is divided into 2 that is with the commutation signal SPWM (Sinusoidal Pulse Width Modulation) in the form of sine wave and signal commutation of half bridge inverter (commutation 180□ & commutation 120□) in the form of square wave. The ideal inverter waveform is a sinusoidal, but in reality, the output of the inverter is not pure sinuoidal and contains a high enough harmonics. For low and medium power, inverter wave output in the form of symmetry or box is still allowed, but for high power required voltage wave approaching sinusoidal and small harmonics.

The three-phase inverter commutation arrangement in this research uses the half bridge inverter (120^0) commutation commutation) signal [3]. The following is a three-phase iverter circuit in Fig. 1.

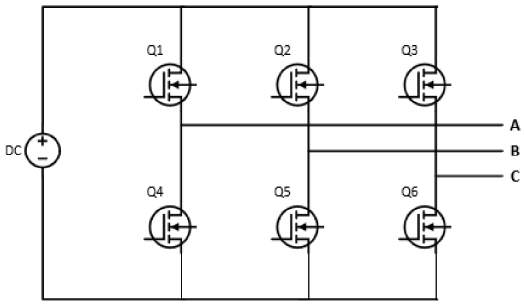


Fig. 1. Three-phase inverter circuit

In Fig. 1. above is a three-phase inverter circuit image that can be used to run BLDC motors. The meaning of the 120^o commutation arrangement is that each switch / mosfet of the inverter works ON for 120^o and OFF for 240^o. Below is the commutation wave signal in the inverter shown in Fig. 2.

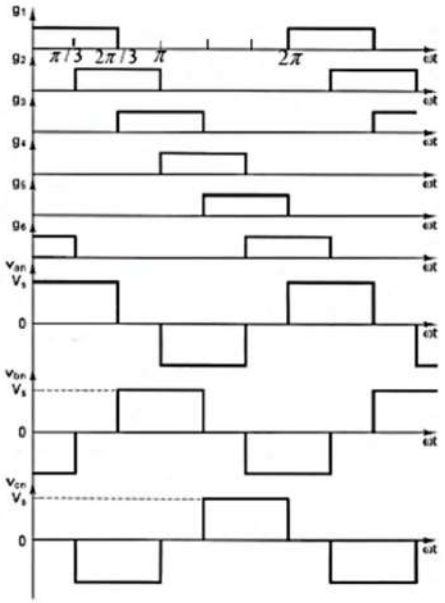


Fig. 2. Implementation of Six-Step Algorithm

Table I Represents the on / off switch table on the inverter for CW rotation of the BLDC motor. The on / off switch of the inverter is determined from the hall input of the sensor on the microcontroller [4] .

TABLE I. ON / OFF INVERTER SWITCH FOR CW MOTOR ROUND BLDC.

Hall Sensors Value (H3 H2 H1)	Phase	Switches
110	A-C	S1 ; S2
100	A-B	S1 ; S6
101	C-B	S5 ; S6
001	C-A	S5 ; S4
011	B-A	S3 ; S4
010	B-C	S3 ; S2

Each switch in a conductive three-phase inverter pairs for 600 in sequence according to the position of the hall effect sensor as shown in Table 2.1. From the table, when the hall effect sensor position 101, Q1, Q5, Q6 conduction, the current from the VDC source flowing through Q1 & Q5 to the AB coil, and return to the source VDC through Q6. At the moment of effect hall effect sensor 100, Q1, Q2, Q6 conduction, current from VDC source flowing through Q1 to A-C coil, and return to VDC source through Q2 & Q6. When the hall effect sensor position 110, then Q1, Q2, Q3 conduction, the current from the VDC source flowing through Q1 & Q3 to the B-C coil, and return to the source VDC through Q2. At the moment of hall effect of sensor 010, then Q2, Q3, Q4 conduction, current from source VDC flowing through Q3 to coil B-A, and return to VDC source through Q2 & Q4. At the moment of hall effect flowing of sensor 011, then Q3, Q4, Q5 conduction, current from source VDC flowing through Q3 & Q5 to coil C-A, and return to VDC source through Q4. At the position of hall effect sensor 001, then Q4, Q5, Q6 conduction, current from source VDC flowing through Q5 to coil C-B, and return to VDC source through Q4 & Q6. Each MOSFET will be conducting / ON for 120□ in each period.

The inverter drive signal in the six-step BLDC motor speed adjustment system using the PWM (Pulse Width Modulation) signal is shown in Fig. 3.

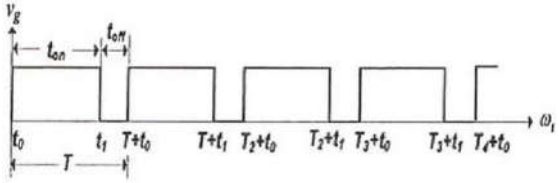


Fig. 3. PWM Signal with Duty Cycle Setup

The PWM signal in Fig. 3, period T is kept constant, while ton and toff are made variation (adjustable). When this signal drives the switch, then the output voltage is:

$$Vo = \frac{1}{T}\int_{t_0}^{T+t_0} V_0(dt) \quad (1)$$

$$Vo = \frac{1}{T}\left\{\int_{t_o}^{t_1} Vo(dt) + \int_{t_1}^{T} 0(dt)\right\} \quad (2)$$

$$Vo = \frac{1}{T}\int_{0}^{t_1} V_0(dt) \quad (3)$$

$$Vo = \frac{t_1 - t_0}{T} V_s \quad (4)$$

$$Vo = \frac{t_{on}}{T} V_s \quad (5)$$

$$Vo = DV_s \quad (6)$$

Where Vo is the output voltage and Vs is the source voltage. T is the wave period and ton is the time on the wave.

B. Brushless DC Motor

DC and BLDC motors have a major difference that is implied by its name. Conventional dc motors have a brush attached to the stator while the DC motor "brushes" does not. Another important part of BLDC is that hall sensor, hall sensors are systematically attached to the rotor and they are used as major sensing devices by Hall Effect sensors embedded into the stator while DC motors have no hall sensors to regulate motor rotation. BLDC motors operate in many modes (phase), but the most common is 3 phases. 3-phase BLDC has better efficiency and provides very low torque. Although, it has some cost implications, the 3-phase has excellent precision in control [5].

The brushless DC motor equivalent circuit can be shown as in Fig. 4.

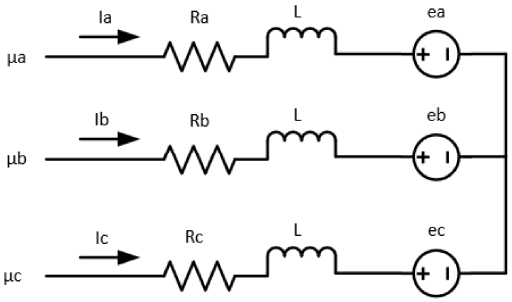


Fig. 4. BLDC Motor Replacement Circuit

The matrix of the brushless DC motor phase voltage equation can be shown as follows:

$$\begin{bmatrix} Ua \\ Ub \\ Uc \end{bmatrix} = \begin{bmatrix} R & 0 & 0 \\ 0 & R & 0 \\ 0 & 0 & R \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \begin{bmatrix} L & 0 & 0 \\ 0 & L & 0 \\ 0 & 0 & L \end{bmatrix} p \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} + \begin{bmatrix} e_a \\ e_b \\ e_c \end{bmatrix} \quad (7)$$

Where, uA, uB, uC are the phase voltages A, B, and C. R is the stator resistance, c. L and M are self inductance and mutual inductance. iA, iB, iC are phase currents A, B, and C. eA, eB, eC are backlashes of phases A, B, and C [6]. Electromagnetic torque of DC brushless motor can be displayed as follows:

$$T_e = \frac{e_A i_A + e_B i_B + e_C i_C}{\omega_m} \quad (8)$$

Where, T_e is the electromagnetic torque of the brushless DC motor and m ω is the angular spin speed in radians per second. The equations for calculating motion for constructing complete mathematical models of electromechanical systems can be shown as follows:

$$T_e - T_L = J\frac{d\omega_m}{dt} + B_v\omega_m \quad (9)$$

Where, TL is the load torque, J is the moment of motor inertia, and BV is the thick coefficient of friction. Electrical position of the rotor and the rotor speed associated with.

$$\frac{d\theta_r}{dt} = \frac{p}{2}\omega m \quad (10)$$

Where, r is the rotor position with respect to the d-axis in rad and P is the number of poles.

II. Soft Starting & Speed Control For Brusless DC Motor

The BLDC motor speed setting in this paper is set using the PWM technique as input from the inverter and varying the DC voltage as input to the DC brushless motor. Motor speed setting can be done by changing the amount of input voltage provided on the motor BLDC in accordance with the desired motor speed. Motor speed setting can also be done by changing the pulse duty cycle. The pulse that varies its duty cycle can determine the speed of the BLDC motor.

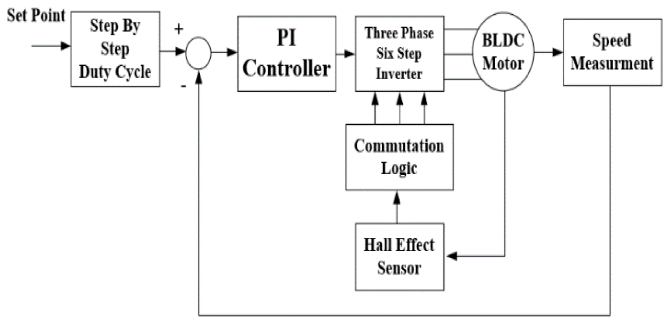


Fig. 5. Block Diagram of Speed Control from Brushless DC Motor

The duty cycle in this paper is generated from comparing 2 control signals is sawtooth with a step type 2 signal [7]. The speed control of the block diagram of the brushless DC motor is shown in Fig. 5.

III. Design Of Controller

The system controller is used to shorten the rise time and return the desired set point value, even if the set point value changes [8]. For the block diagram of the PI controller can be shown by Fig. 6.

:

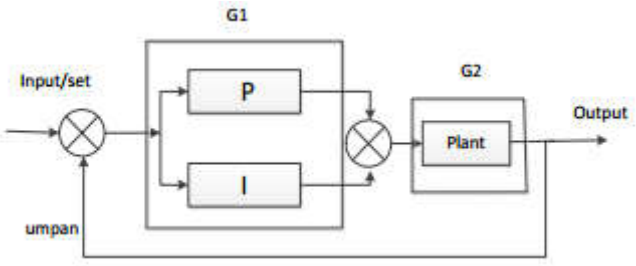


Fig. 6. Block Diagram of PI controller

Based on to the block diagram of the PI control, the following mathematical calculation will be performed:

$$G2 = \frac{C(S)}{K(S)} = \frac{K}{\tau s+1} \quad (11)$$

$$G1 = Kp\left[\frac{\tau is+1}{\tau is}\right] \quad (12)$$

Equation of OLTF (*open loop transfer function)*

$$\frac{C(S)}{K(S)} = \frac{K}{\tau s+1} \quad (13)$$

Equation of CLTF (*close loop transfer function)*

$$\frac{C(S)}{R(S)} = \frac{Kpk\frac{(\tau is+1)}{\tau is}.\frac{1}{(\tau s+1)}}{1+Kpk\frac{(\tau is+1)}{\tau is}.\frac{1}{(\tau s+1)}} \quad (14)$$

$$\tau i = \tau \quad (15)$$

$$\frac{C(S)}{R(S)} = \frac{\frac{Kpk}{\tau is}}{1+\frac{Kpk}{\tau is}} = \frac{1}{\frac{\tau is}{1+Kpk}+1} \quad (16)$$

$$\frac{C(S)}{R(S)} = \frac{1}{\frac{\tau i}{Kpk}s+1} \quad (17)$$

$$\frac{C(S)}{R(S)} = \frac{1}{\tau^* s+1} \quad (18)$$

IV. SIMULATION & ANALYSIS RESULT

To validate the application of the soft starting and PI Speed Controller, the simulation is done on BLDC motor in PSIM. This simulation evaluates the performance of the starting & PI control on the BLDC motor with a normal starting BLDC motor. to compare the starting current when using the soft starting. Fig. 7 represents the BLDCM simulation model for soft starting with PI speed controller in PSIM.

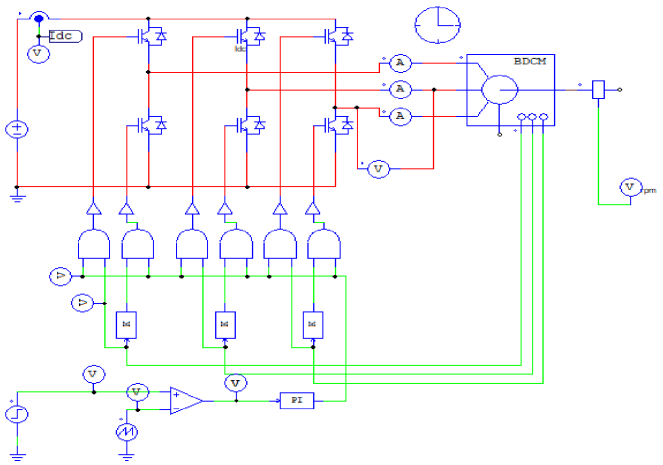


Fig. 7. Simulation Series On PSIM

The BLDC motor parameters used in this simulation are given in Table II.

TABLE II. BLDC MOTOR PARAMETERS

Voltage Input	48V
Phase Resistance	11.9Ω
Phase Inductance	0.00207H
Amount of poles	4
Rated speed	1000Rpm

The duty cycle setting value in this simulation is 90%. The duty cycle will be upgraded by step by step by comparing the step type 2 signal with saw tooth. Where the comparison output of two signals will result in step duty adjustment that can be adjusted according to the time set. In this paper 90% duty cycle will be achieved with 10ms time. Fig. 8 shows the duty cycle wave with soft starting BLDC motor.

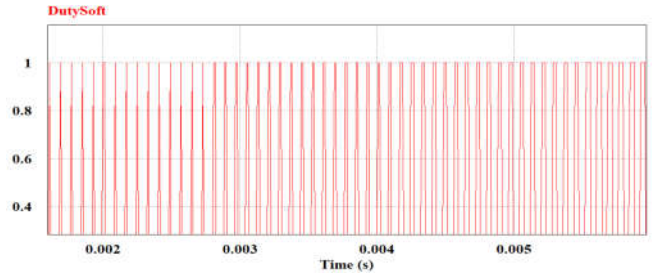


Fig. 8. Duty Cycle Soft Starting On PSIM

Fig. 9 shows duty cycle waves with normal starting BLDC motors.

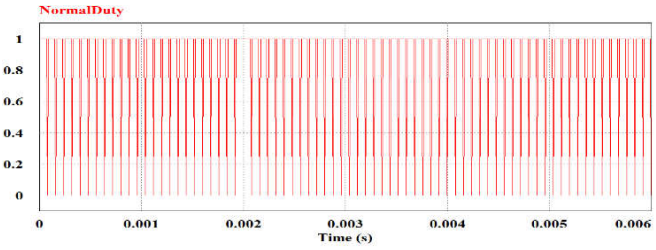


Fig. 9. Duty Cycle Normal On PSIM

Fig. 10 shows the response of BLDC motor characteristics when using soft starting method. On the current curve shows the steady current value of 0.389A and the over-shoot current reaches 0.74A.

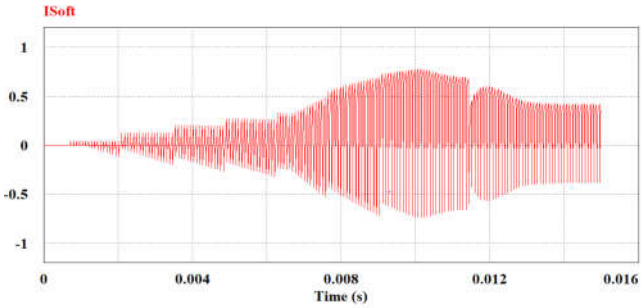


Fig. 10. Input Current Soft Starting In PSIM

Fig. 11 shows the response of BLDC motor characteristics when using normal starting method. On the current curve shows the steady current value of 0.54A and the over-shoot current reaches 1.3A.

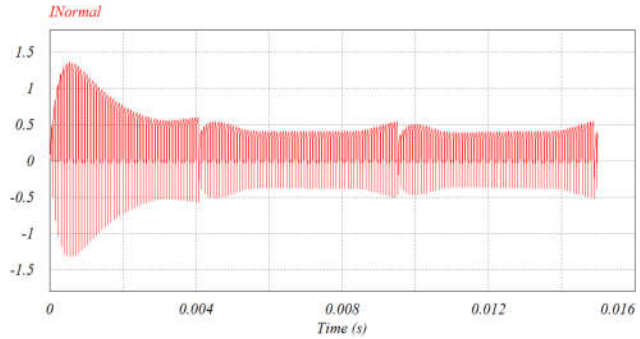


Fig. 11. Current Input Normal Starting On PSIM

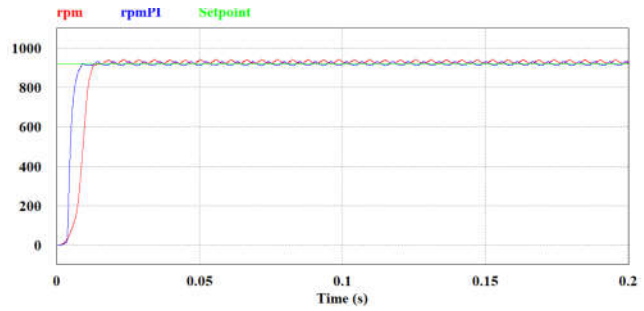


Fig. 12. Current Input Normal Starting On PSIM

Fig. 12. shows the response of BLDC motor characteristics when using PI controller method and without PI controller. The BLDC set point in this simulation is 900Rpm. At the speed curve shows the speed reaching the set point when using the PI control is 0.0088s. whereas At the speed curve shows the speed reaches the set point when without using PI control is 0.0138s.

V. CONCLUSION

BLDC motor soft start simulation has been presented. The simulation results show that the soft starting method that is tuned by comparing the step type 2 signal with the saw toot

signal can decrease over shoot of BLDC starting motor current. Soft starting indicates that the starting current on the BLDC motor is smaller than the normal starting. In addition, the use of PI controller in this paper is proven to shorten BLDC motors reaching the set point motor rpm specified. This research is the first step of implementation. So, in the future this research will be a reference to be continued in implementation.

REFERENCES

[1] Jaya A., Purwanto E., Fauziah M.B, Murdianto F.D, Prabowo G, Rusli M.R, "Design of PID-Fuzzy for Speed Control of Brushless DC Motor in Dynamic Electric Vehicle to Improve Steady-State Performance" (IES-ETA) International Electronics Symposium on Engineering Technology and Applications, Surabaya, 2017.

[2] Jaya A, Wahjono E, Rusli M.R, Purwanto E, Murdianto F.D, Fauziah M.B, "Design and Simulation of Sensorless BLDC Motor Drive Using Flux Linkage Increment Based on the line-to-line BEMF for Electric Vehicles" (IES-ETA) International Electronics Symposium on Engineering Technology and Applications, Surabaya, 2017.

[3] T.Kenjo, S.Nagamori "Permanent Magnet and Brushless DC Motors".

[4] Mr. Namdev K. Barhate 1 Prof. K. S. Ingle, "Design and Development of Digital PID Controller for Brushless DC Motor Drive System" International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) Volume 5, Issue 3, March 2016.

[5] Issa Batarseh, Power Electronic Circuits. 2004.

[6] Simulating brushless dc motors. Dspace 2009.

[7] Bapayya Naidu Kommula, Venkata Reddy Kota, "Performance Evaluation of Hybrid Fuzzy PI Speed Controller for Brushless DC Motor for Electric Vehicle Application" 2015 Conference on Power, Control, Communication and Computational Technologies for Sustainable Growth (PCCCTSG) December11-12, 2015, Kurnool, Andhra Pradesh, Indi

[8] R.R.Navatakke1, Jagadeesh.Bichagatti.2,"Optimal PID control of a brushless DC motor using PSOtechnique" IOSR Journal of Electrical and Electronics Engineering (IOSR-JEEE) e-ISSN: 2278-1676,p-ISSN: 2320-3331, Volume 10, Issue 3 Ver. IV (May – Jun. 2015), PP 13-17

[9] Murdianto F.D, P. Ontoseno, P. Ardyono, "Modeling and simulation of mppt-bidirectional using adaptive neuro fuzzy inference system (ANFIS) in distributed energy generation system," IEEE International Seminar on Intelligent Technology and Its Applications (ISITIA), 2015.

[10] Soedibyo, Murdianto F.D, Suyanto, A. Mochamad, P. Ontoseno,"Modeling and simulation of mppt SEPIC combined bidirectional control invers KY converter using ANFIS in microgrid system," Indonesian Journal of Electrical Engineering and Computer Science, vol. 1, pp. 264-272, February 2016.

[11] Suyanto, Murdianto F.D, P. Ontoseno, S. Adi, S. Rony,"Wind-PV hybrid system modeling using bidirectional converter with MPPT-dual adaptive neuro fuzzy inference system (ANFIS) in microgrid isolated system," Journal of Engineering and Applied Sciences, vol. 11, pp. 2353-2359, December 2016.

[12] Rakhmawati R, Wahjono E, Suhariningsih, Murdianto F.D, Andriawan D,"Design and Implementation Boost Converter with Constant Voltage in Dynamic Load," International Seminar on Application for Technology of Information and Communication (iSemantic), 2017

[13] Efendi M.Z, Murdianto F.D, Setiawan R.E,"Modeling and simulation of MPPT sepie converter using modified PSO to overcome partial shading impact on DC microgrid system," International Electronics Symposium on Engineering Technology and Applications (IES-ETA), 2017.

[14] Murdianto F.D, Nansur A.R, Hermawan A.S.L,"Modeling and simulation of MPPT coupled inductor sepie converter using Flower Pollination Algorithm (FPA) Method in DC mierogrid system," International Electronics Symposium on Engineering Technology and Applications (IES-ETA), 2017.

[15] F.D. Murdianto, M. Z. Efendi, R. E. Setiawan and A. S.. L. Hermawan, "Comparison method of MPSO, FPA, and GWO algorithm in MPPT SEPIC converter under dynamic partial shading condition", International Conference on Advanced Mechatronics, Intelligent Manufacture, and Industrial Automation (ICAMIMIA), 2017

[16] F.D. Murdianto, A. S.. L. Hermawan, A. R. Nansur and R. E. Setiawan, "Comparison method of flower pollination algorithm, modified particle swarm optimization and perturb & observe in MPPT coupled inductor sepic converter on DC microgrid isolated system", International Conference on Advanced Mechatronics, Intelligent Manufacture, and Industrial Automation (ICAMIMIA), 2017

Performance Robustness of PID Controller in Buck Converter For Cooling System

Renny Rakhmawati
Departement Electrical Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
renny@pens.ac.id

Janitra Hilmyvarafi Farrasbyan
Departement Electrical Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
farrasbyan@gmail.com

Suhariningsih
Departement Electrical Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
suhariningsih@pens.ac.id

Farid Dwi Murdianto
Departement Electrical Engineering
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
farid@pens.ac.id, fariddm@ieee.org

***Abstract*— When driving in the long journey to store the beverages required cooling system that can keep drinks stay fresh and the quality is maintained when it is about to be consumed. However, many cooling systems are now made using compressor and freon systems which use chemicals that are not environmentally friendly, and consume a lot of energy from car batteries. So that the need for other sources of energy that its use will never run out, that is by the use of solar energy with solar cells. However, the output voltage generated by solar cells is always changing based on the intensity of light and weather conditions. Need a control that can adjust the output voltage of solar cells to be stable, if there is any change of output voltage then the controller will be active and quickly restore the voltage in a stable state. Therefore, this paper presents a cooling system that uses solar cells as an energy source which later output voltage will be derived using a DC-DC Buck Converter. Then the output voltage of this buck converter is controlled by the PID control which will be fed back by the voltage sensor for battery charging process. This battery will be an energized source for Peltier as a cooling device in the cooling beverage system. It is hoped that from this paper can be designed and implemented environmentally friendly cooling system and utilize renewable energy and save energy on car battery.**

Keywords—Peltier, Solar cells, DC-DC Buck Converter,PID Control

I. INTRODUCTION

Cooling systems are generally widely used in electronic appliances such as refrigerators, dispensers, and others. And in its development is also widely used to engine cooling system. Cooling systems in circulation today generally use a compressor system and Freon which uses chemicals that are not as environmentally friendly as Freon gas. This should be an important concern because Freon gas can damage the ozone layer, which is endangering the lives of mankind.

Therefore, new alternatives / breakthroughs are introduced to make refrigerators easier to manufacture and not to use chemicals so environmentally friendly, is the use of electrical cooling modules (peltier) in the cooling system aims to lower the temperatures in the cooling chamber, in addition to peltier also more environmentally friendly because peltier does not contain Freon gas. But electrical cooling (peltier) has a weakness that requires considerable power consumption.

Basically the cooler beverage box that is made now still use the battery in the vehicle for the source of electrical energy and there is no temperature setting inside. So it's very needed the temperature setting to keep the freshness of the beverage so that when consumed still fresh. This beverage cooling system, aimed at people who are traveling on a long trip and take a long time, so there will be no trouble the needs of fresh beverage.

II. TERMOELEKTRIK EFFECT & THERMOELECTRIC COOLER

A. Thermoelectric Effect.

Before the thermoelectric effect is found, T. J. Seebeck discovers a seebeck effect in which an electric motion force (emf) can be generated by heating the connection points between two different electrical conductors. Seebeck effect is the embryo of peltier effect. The Seebeck effect can be demonstrated by making the connection between two wires of different metal types then the other end of the wire connected to the galvanometer. If the connection between the wires is heated, the measuring device will read a small amount of voltage as shown in Figure 1. Two wires can be said to form a thermocouple. It is also known that the thermoelectric voltage is proportional to the temperature difference between the thermocouple connection point and its connection to the measuring device.[9]

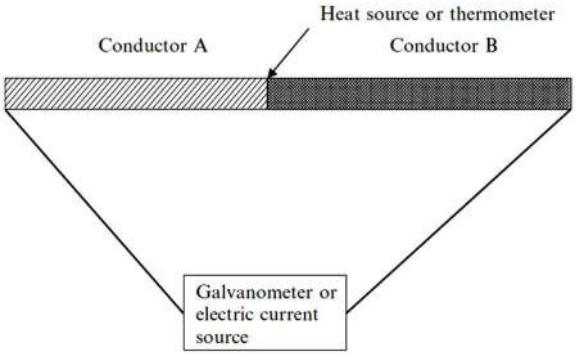


Fig. 1. Experiments showing Seebeck and Peltier effects

Peltier effect can be demonstrated, in principle, replacing the meter with a direct current source and forging a small thermometer at the junction point of the thermocouple.[7][15]

B. Thermoelectric Cooler

Thermoelectric cooler is a device that can cause a temperature difference between the two sides if a direct electric current flows in both poles, in this case the semiconductor. Figure 2 shows the physical shape of the peltier element. In terms of refrigeration, peltier elements have the advantage of absence of vibration or fluid use, and their small size and shape are easily engineered. While the deficiency lie in the factor of low power efficiency and cost system design is still relatively expensive. However, now many researchers are trying to develop efficient and cheap peltier elements.

Figure 3 shows a peltier element composed of a series of two types of semiconductors (tip e-p and n-type) connected in series. Each connection between the two types of semiconductors is connected to a conductor made of copper. The interconnect of the conductor is placed respectively at the top and bottom of the semiconductor. The top conductor is intended to remove the heat and the lower conductor is intended to absorb the heat. On both sides of the interconnect are affixed plates made of ceramic. This plate aims to concentrate the heat derived from the conductor.[7][14]



Fig. 2. The physical shape of the peltier element

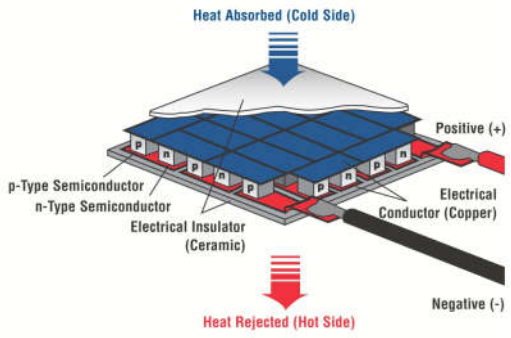


Fig. 3. TEC module composition

Figure 4 shows a peltier of an electric current element and causes a temperature difference in the two interconnections. The current interconnected from the n-type to the p-type semiconductor will absorb the heat or in other words become cold. Meanwhile, the current-flown interconnection from the p-type to the n-type semiconductor will dissipate the heat or in other words become hot. The interconnections between semiconductors on peltier elements are made of conductors that cause currents to flow in both directions.[9][15]

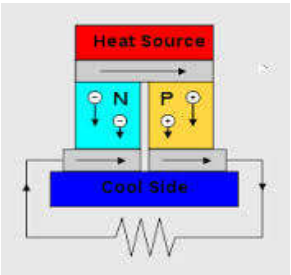


Fig. 4. Peltier element Workflow

Figures 1 through 4 above are quoted from references[9][12][14].

III. MODELING OF BEVERAGEY COOLING SYSTEM

In this chapter explained about the design of drinking cooling system such as block diagram, and design of buck converter.

A. Block Diagram Beverage Cooling System

In Figure 1 presents a block diagram of a beverage cooling system with an energy source from a solar panel that processes solar energy into a DC voltage source[16-23]. The DC voltage of the solar panel is then used to charge the battery before using it as a supply for peltier. The output voltage of the solar panel will be read by the V1 sensor then the ARMSTM32F4 microcontroller will sense as PID control[24-25]. That will be the basis of the duty cycle value of PWM. Sensors v2 and i1 are used to monitor buck converter voltage and current output to remain constant to charge the battery. To charge the battery, it takes a constant voltage of 14.4V. Before the battery is used

to supply the peltier it is first passed on the converter buck to set the right current for the peltier so that the resulting temperature is stable. There is a microcontroller that uses ARM STM32F4 tasked to set all controls work as data conversion current and voltage sensors, as well as for control of PWM buck converter.[1][6]

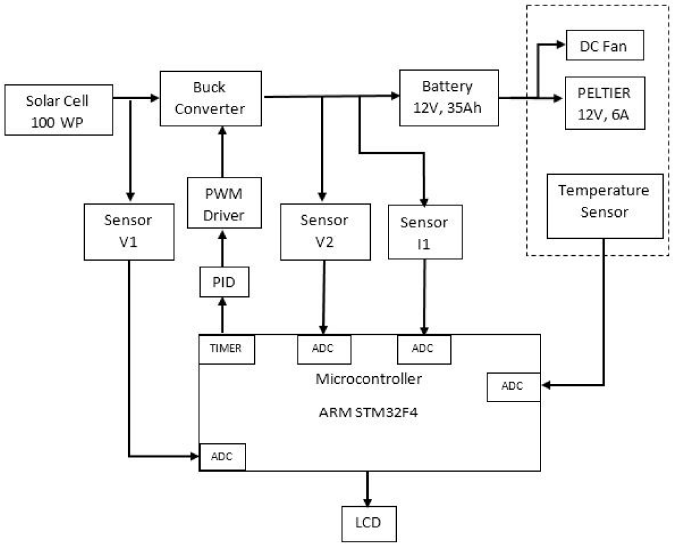


Fig. 5. Block diagram of the beverage cooling system

as shown in figure 5 above the PID control settings will be set using the PWM (Pulse Width Modulation) method. The microcontroller functions to adjust the pwm value which will then be displayed by LCD (Liquid Cristal Display).

B. Design Buck Converter

Buck Converter operates when the switch (mosfet) is closed, the diode becomes reverse-biased and the current flows from the source to the inductor, filtered by the capacitor then towards the load, and returns to the source. When the switch (mosfet) is opened, the diode becomes foward-biased and carries the inductor current. Then the current stored in the inductor will be released to the load [1,4,5].

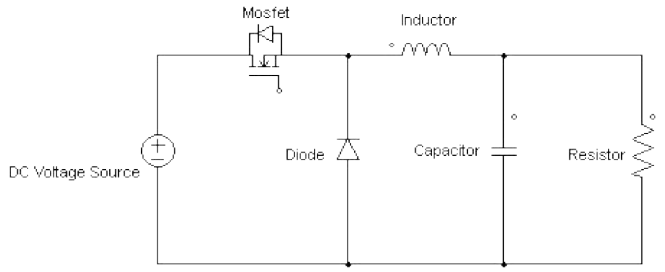


Fig. 6. Buck converter circuit

$$V_o = VsD \quad (1)$$

$$L_{\min} = \frac{(1-D)R}{2f} \quad (2)$$

$$L = 10xL_{\min} \quad (3)$$

$$C = \frac{1-D}{8L(\Delta V_o / V_o)f^2} \quad (4)$$

The meaning of the symbol above is as follows:

Vo : output voltage
Vs : input voltage
D : duty cycle
R : resistor
L : inductor
C : capacitor

Equation 1 (output voltage), 2 (minimum inductor value), 3 (inductor value), and 4 (capacitor value) are required to design buck converter. In this paper buck converter is used for battery charging with constant voltage control 14.4V. Table 1 presents the design value of each component to improve the converter and figure 7 hardware of the buck converter.[4][5]

TABLE I. DATA SPESIFICATION OF BUCK CONVERTER

Parameter	Value
Switching frequency (fs)	40kHz
Input Voltage (Vin)	12V
Output Voltage (Vout)	14.4V
Inductor (L)	353uH
Capacitor (C)	22uF
Resistor (R)	4.1Ω

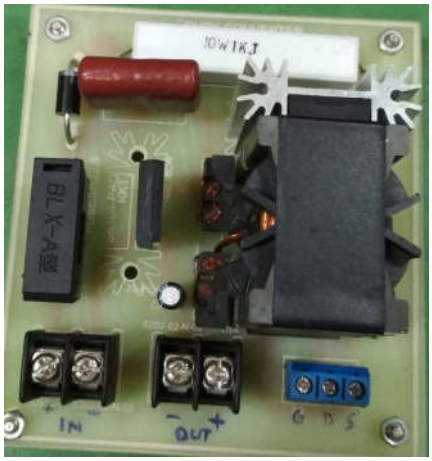


Fig. 7. Buck converter

IV. Simulation and Experimental Results

In this chapter described the comparison between simulation in Simulink matlab with the result of measuring output voltage on buck converter as well as testing the temperature of storage room in beverage cooling system.

A. Buck Converter with Constant Voltage Control Simulation

Figure 8 presents the matlab simulink block at input available reference voltage input, PID control, and transfer function of the buck converter. Figure 9 shows the output response voltage of the buck converter when open loop or without control while Figure 10 shows the output voltage response of the controlled buck converter with the parameter values as shown in the design.[1-6]

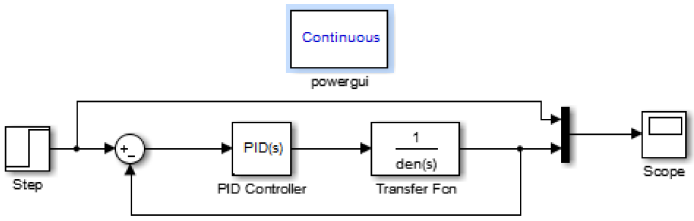


Fig. 2. Simulation of buck converter with fixed voltage control

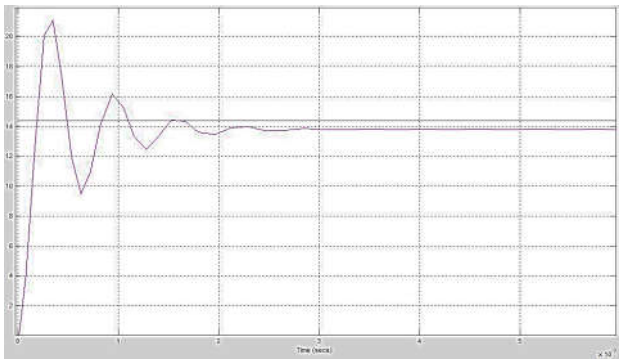


Fig. 3. Output voltage response buck converter without control

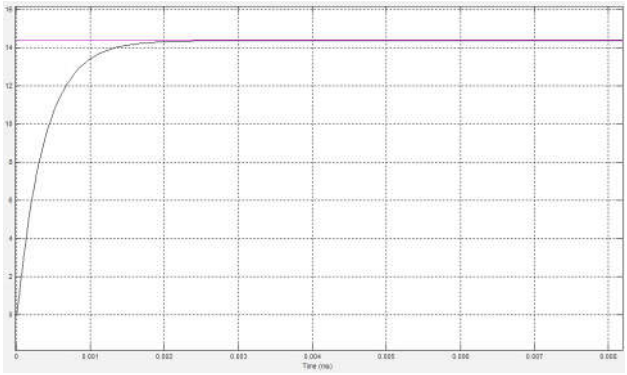


Fig. 4. Output voltage response buck converter with control

B. Buck Converter Experimental

The purpose of this buck converter experiment is to know the output voltage of buck converter if given PWM input with fixed duty cycle and compare it with design voltage so that it is known how much error between the measurement and theory so it is known whether the buck converter works well or not. [6]

TABLE II. EXPERIMENTAL RESULT OF BUCK CONVERTER

Duty Cycle	Vin (V)	Iin (A)	Vout (V)	Iout (A)
0.2	10	0.07	1.44	0.4
	15	0.14	2.363	0.65
	20	0.2	3.3	0.9
	22	0.23	3.677	1
	25	0.26	4.23	1.5
0.4	10	0.36	3.244	0.97
	15	0.58	5.07	1.4
	20	0.77	6.86	1.95
	22	0.8	7.61	2.18
	25	1	8.64	2.4
0.5	10	0.59	4.16	1.2
	15	0.9	6.41	1.8
	20	1.2	8.69	2.4
	22	1.37	9.6	2.7
	25	1.5	10.96	3
0.65	10	1	5.53	1.6
	15	1.5	8.47	2.4
	20	2	11.34	3.18
	22	2.23	12.53	3.5
	25	2.58	14.2	4

Table 3 shows the results of buck converter experiment with different duty cycle that is 0.2,0.4,0.5, and 0.65 with input voltage that varies in each duty cycle. The load used in the experiment is 4.1 Ω and the sampling frequency is 40 kHz. Based on the results obtained in Table 4, the converter is good enough with a duty cycle of 0.65 with the largest output voltage of 14.2 which is close to the design of 14.4 V with 1.3% error.

C. Storage room temperature experimental with battery

This experiment aims to determine the current consumed by the peltier when cooling the coolant box temperature with reference to the peltier power using a voltage source of the 12V, 35Ah battery. Peltier testing in this cooler box is used arrangement 3 peltier which is installed in parallel and in condition without load in room still empty room. The following is the result of the test of temperature shown in Table 3. The result of cold temperature picking of peltier. As

well as in figure 11 is also shown the response of room temperature to time.[8][15]

TABLE III. STORAGE ROOM DATA

t (minute)	Temperature (°c)
0	29.1
10	27.1
20	26.7
30	24.8
40	22.6
50	20.9
60	18.6
70	17.2
80	15.8
90	13.3
100	12.7
110	11.5

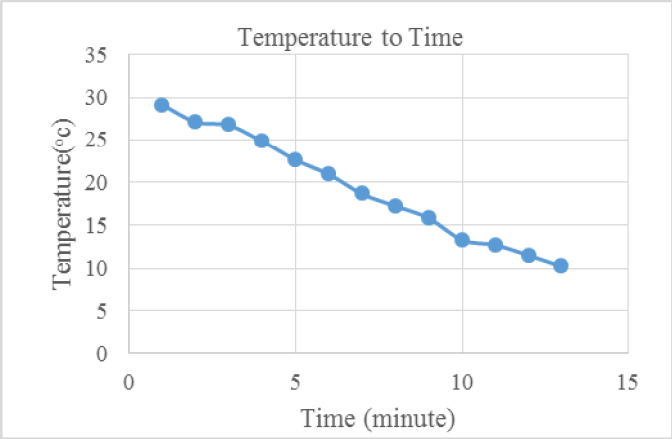


Fig. 11. Graph of temperature testing of the room

Based on the test data in Table 3 the lowest temperature produced by 3 peltiers arranged in parallel with a 12V voltage of 11.5 °C. With a power consumption of 97.5 W as well as a fairly large current 8.2A. Can be observed from the experimental results to cool box size 325 x225 x 285 mm with a capacity of 12 liters. It takes a long time of 110 minutes to achive cool box temperature 11.5 °C.

D. Storage room temperature experimental with voltage change

This test aims to find out how the ideal voltage and current required from peltier to cool the room temperature at the beverage storage. As well as find out the temperature response to the time before it is controlled and in no-load conditions in the room is still empty room. In this system peltier given voltage source from power supply between 6-7 volt, with 3 peltier which arranged parallel. Here is the result of the room temperature response to time, data retrieval is done 10 minutes once shown in table 4.[13][14]

TABLE IV. STORAGE ROOM DATA

Temperature (°c)	Vin (V)	Iin (A)	t (minute)
29.6	6.16	3.9	0
26.9	6.24	4	10
25.5	6.5	4.3	20
24.5	6.7	4	30
21.8	6.78	4	40
20.8	6.8	4	50
19.9	6.9	4.4	60
19.2	6.9	4.4	70
18.7	7	4.5	80
18.4	7.03	4.5	90

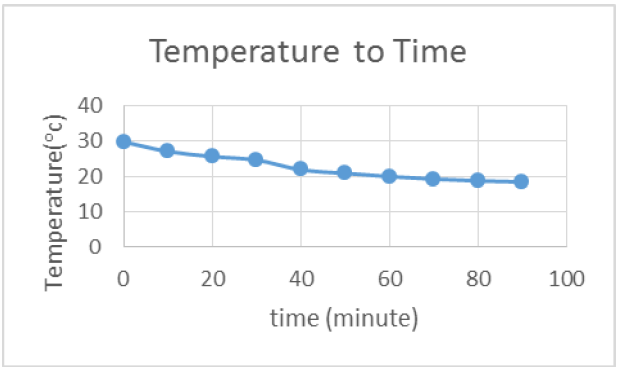


Fig. 12. Graph of temperature testing of the room

It can be seen from the test data in Table 4 that the room temperature generated by three peltiers arranged in parallel is cooler at 18.4 °C. This explains that with the addition of peltier use can accelerate the cooling time of room temperature for 80 minutes. Also applies to lower power consumption and lower source currents with an average power consumption of 27.5W and average current consumption of 4.5A.

V. CONCLUSION

Based on the simulation results above it can be seen that the PID control can work well reducing voltage overshoot and accelerating rise time and reducing oscillation. By using the control parameter value kp = 1.5, ki = 200, kd = τd = 0.0007143, and τi = 27664s. This battery charger system can work in a period of 3 hours and produces a 13.8 Volt charging voltage and the results of the battery charger are functioned as peltier load as coolant drink room. With the achievement of temperature in the cooling room reaches 11.5 ° C.Based on the results of peltier testing that has been carried out, the maximum room temperature reaches 11.5°C with a peltier

consumption voltage of 11.59volt and a current of 8.9A. With 3 peltiers arranged in parallel. The main source of supply is obtained from 100 WP solar cell energy. This research is the first step of implementation. So, in the future this research will be a reference to be continued in implementation.

References

[1] A.M. Wu, J. Xiao, D. Markovic, S.R. Sanders, "Digital PWM control: Application in voltage regulator models", *IEEE Power Electronics Specialists Conference 1997 Record*, pp. 77-83.

[2] K. Vibhute and R. Shukla, "All weather solar cell- a new trend of design of solar cell",*Electrical Power and Energy Systems (ICEPES), International Conference on*, Bhopal, 2016.

[3] M. H. Rashid, "Power Electronics for Renewable", *Power Electronics Handbook*, USA, Elsevier's Science & Technology Rights Department, 2011, p. 723.

[4] T.W. Martin, S.S. Ang, "Digital control of switching converters", *IEEE Symposium on Industrial Electronics*, vol. 2, pp. 480-484, 1995.

[5] D. W. Hart, "The Buck Converter", *Power Electronics*, New York, McGraw-Hill, 2011, pp. 211-220.

[6] R. R. Boudreaux, R. M. Nelms, J. Y. Hung, L. C. Mathison, "Digital control of a buck converter using an 8-bit microcontroller", *Technical Papers of the Tenth International High Frequency Power Conversion 1995 Conference*, pp. 238-251, 1995.

[7] Zhang HY. A general approach in evaluating and optimizing thermoelectric coolers, International Journal of Refrigeration 2010, 1187:96-33(6).

[8] Rashid, Muhammad H rashid,"Power Electronics Circuits, Devices, and Application 2nd Ed, Jakarta, PT Prenhallindo, 1999.

[9] Dongliang Zhao & Gang Tan. *A Review of Thermoelectric Cooling:Material, Modeling and Applications*. Applied ThermalEngineering 2014, Vol. 66, pp 14-24.

[10] Zhou, Y. and Yu, J. Design optimization of thermoelectric cooling systems for applications in electronic devices. *International Journal of Refrigeration* 2011, **35**:1139-1144.

[11] Moin Ahmad S. Shaikh, M.K. Chopra, *An Extensive Review on Thermoelectric Refrigerator*, International Journal of Scientific Progress and Research 2014, Vol. 6 no. 1, pp 7-11.

[12] R. Umboh, J.O. Wuwung, E. Kendek Allo, B.S. Narasiang, , Perancangan Alat Pendinginan Portable Menggunakan Elemen Peltier, Vol. 1 no. 3, e-Journal Teknik Elektro dan Komputer,2012 pp 1-6.

[13] Yang D & Yin H. Energy Conversion Efficiency of a Novel Hybrid Solar System for Photovoltaic, Thermoelectric, and Heat Utilization, IEEE Trans. Energy Conv. 2011; 26(2): 662:70-26(2).

[14] H. J.Goldsmid, *Introduction to Thermoelectricity*, London, 2009.

[15] Palash Nakhate, Niraj Pawskar, Purva vatamwar, Saurabh Kalambe. Ecofreindly Refrigerator Using Peltier, International Journal of Research In Science & Engineering (IJRISE), Special Issue (7): 2017

[16]F.D. Murdianto, M. Z. Efendi, R. E. Setiawan and A. S.. L. Hermawan, "Comparison method of MPSO, FPA, and GWO algorithm in MPPT SEPIC converter under dynamic partial shading condition", International Conference on Advanced Mechatronics, Intelligent Manufacture, and Industrial Automation (ICAMIMIA), 2017

[17]F.D. Murdianto, A. S.. L. Hermawan, A. R. Nansur and R. E. Setiawan, "Comparison method of flower pollination algorithm, modified particle swarm optimization and perturb & observe in MPPT coupled inductor sepic converter on DC microgrid isolated system", International Conference on Advanced Mechatronics, Intelligent Manufacture, and Industrial Automation (ICAMIMIA), 2017

[19]F. D. Murdianto, O. Penangsang, and A. Priyadi, "Modeling and simulation of mppt-bidirectional using adaptive neuro fuzzy inference system (ANFIS) in distributed energy generation system," IEEE International Seminar on Intelligent Technology and Its Applications (ISITIA), 2015.

[20]M. Z. Efendi, F.D. Murdianto and R. E. Setiawan, "Modeling and Simulastion of MPPT Sepic Converter using Modified PSO to Overcome Partial Shading Impact on DC Microgrid System," International Electronics Symposium on Engineering Technology and Applications (IES-ETA), 2017

[21]F. D. Murdianto, A. R. Nansur, and A. S.. L. Hermawan, "Modeling and Simulation of MPPT Couple Inductor Sepic Converter using Flower Pollination Algoritm (FPA) Method in DC Microgrid," International Electronics Symposium on Engineering Technology and Applications (IES-ETA), 2017

[22]Soedibyo, Murdianto F.D, Suyanto, A. Mochamad, P. Ontoseno,"Modeling and simulation of mppt SEPIC combined bidirectional control invers KY converter using ANFIS in microgrid system," Indonesian Journal of Electrical Engineering and Computer Science, vol. 1, pp. 264-272, February 2016.

[23]Suyanto, Murdianto F.D, P. Ontoseno, S. Adi, S. Rony,"Wind-PV hybrid system modeling using bidirectional converter with MPPT-dual adaptive neuro fuzzy inference system (ANFIS) in microgrid isolated system," Journal of Engineering and Applied Sciences, vol. 11, pp. 2353-2359, December 2016.

[24Jaya A., Purwanto E., Fauziah M.B, Murdianto F.D, Prabowo G, Rusli M.R, "Design of PID-Fuzzy for Speed Control of Brushless DC Motor in Dynamic Electric Vehicle to Improve Steady-State Performance" (IES-ETA) International Electronics Symposium on Engineering Technology and Applications, Surabaya, 2017.

[25]Rakhmawati R, Wahjono E, Suhariningsih, Murdianto F.D, Andriawan D,"Design and Implementation Boost Converter with Constant Voltage in Dynamic Load," International Seminar on Application for Technology of Information and Communication (iSemantic), 2017

Analysis of 4G Network and Chat Applications to Smartphone Battery Life

Andik Setyono, Dewa Arga Nisadha, De Rosal Ignatius Moses Setiadi
Department Informatics Engineering
Dian Nuswantoro University
Semarang, Indonesia
Email : andik.setyono@dsn.dinus.ac.id, dewaituaku@gmail.com, moses@dsn.dinus.ac.id

***Abstract*— this research aims to prove about the public perception of spendthrift of a smartphone battery life. Many people assume that their smartphone include wasteful with certain types of the use. In this study, the testing subject is the use of smartphones with a standard battery capacity ranging from 1500-2000 mAh, as well as the parameters of chat applications. Determination of the subject is done by establishing the list of event variables, which consists of 216 events contained in these parameters. The data was collected by doing observation directly in the field. Then, the collected data was processed and analyzed statistically by using the classification method with C4.5 algorithm. We also compared the results of manual calculations with WEKA tool and a simple application to test and evaluate the performance of the algorithm. The results showed that, the performance on manual calculation and simple application simulation generated a maximum accuracy of 100%. While, the WEKA get 93.06% accuracy value and experience a little error because there are some additional conditions performed on the calculation on the WEKA. This research contributes real to the science and add insight to the public that will have a true perception about the battery life of the smartphone.**

Keywords*— *Battery Life; Chat Application; 4G Network; Decision tree; C4.5 Algorithm

I. INTRODUCTION

The needs of the people regarding to browsing, chatting and even online games require a sufficient and stable internet network connection. The more often of the smartphone is connected to the internet then the battery power consumption will be more wasteful, because the smartphone antenna needs battery power to capture the signal of a cellular network. The use of internet on a smartphone depends on the used cellular network operators. In Indonesia, 4G network technology is already exist where the spread of the 4G network is not prevalent yet that cause unstable of the 4G network in some areas. The instability in the network certainly has an impact on the performance of the smartphone antenna, which will work harder [1].

One of the factors that influence the strength of the cellular network is the distance between Base Transceiver Station (BTS) to smartphone, weather and external factors such as electrical outage and the user location. Based on these factors, other information will be appeared that can be a great reference for network strength that is the number of signal bars on the smartphone [2], [3]. In addition to the strength of cellular networks, the number of applications that are used also affect the battery life of the smartphone, for example chat applications. Several popular chat applications nowadays are Whatsapp, LINE and BBM. By relying on smartphones that connect to the internet, the users can access chat applications anywhere and anytime. These chat applications have a fantastic number of PlayStore downloaders, BBM with 100 million downloaders, LINE with 500 million downloaders, and Whatsapp with 1 billion downloaders [4].

Initial hypothesis, more and more chat applications that are used then will impact on battery life of smartphone. Besides, chat applications are connected to the internet, there is also an outflow of data. The used data types are in multimedia such as text (chat), image, voice (call) and video (streaming), and also some promotional ads that appear [4]. This makes the data is quite large and certainly also make the smartphone antenna work harder, so that the battery life will be reduced. This hypothesis encourages authors to analyze the impact of smartphone battery life based on the 4G network strength parameters and the number of chat applications used. The generated information are expected to be able to provide insight to the people about the impact of the strength of the available 4G networks and chat applications that are used on smartphone to battery life, so people can be wiser in using smartphones and become more intelligent people.

In this study, we conducted a direct observation to obtain the real data about the network condition in certain areas with reference to the strength of 4G network. Several experiments are performed by using smartphone, then installing the chat applications and adjusted to the condition of the parameters to be searched. After obtaining the required data, then summarized it into a table for easy understanding. The classification method using decision tree C4.5 algorithm is applied as the reference of the data and the creation of the rule manually. In several case studies, the data combination are used to solve existing problems. Data mining is required to process the history data or basis data to be dataset used. The additional data are taken by actual interview or direct observation [5]. The resulted dataset are analyzed by doing simulations use tools (WEKA and developed simple application) as a comparison of the calculation results manually. The generated output will be in the form of information about the battery life of the smartphone with a variety of parameters provided [6].

In the previous study [7], discussed about the effect analysis of 3G signals on smartphones that cause battery fast drop or lowbatt. Since smartphones are already widely circulated and

3G network technology comes in, people are using smartphones as a tool for browsing, chatting and some other media. Browsing and chatting applications require the speed of data packets to be comfortable when in use. However, there is one problem that has been experienced by the user when using his smartphone for browsing and chatting that is battery fast drop or lowbatt. This matter underlies the authors to analyze the impact of 4G network strength and applications that are used to battery life on the smartphone.

The purposes of this research are: (1) to know the implementation of the classification method using decision tree C4.5 algorithm to classify the battery life of smartphone, (2) to find out the battery life of smartphone with parameters the strength of 4G cellular network and the number of chat application used to battery life of smartphone, and 3) to know the performance comparison among manual calculation with result of the computation simulation and the developed simple application.

II. Research Methodology

A. Data Collection Method

Data collection methods as supporting the accuracy and quality of the data that will be used for this research are by conducting survey with direct observation and through literature study. The survey was conducted by observing 4G network to get data of the strength of 4G network based on different conditions and located around Semarang city. Data from chat applications parameters are added on the smartphone used. Surveys for some additional variables such as weather, location and network stability are also performed. The data will be the reference for the research process. Literature review aims to find the source of reference research derived from books, journals, or internet media associated with the conducted research.

B. Data Analysis Method

In this research, the used data analysis method is statistics technique. This is a data analysis method by collecting data or facts which then processed into data more concise, presented in the form of data tables, and analyzed the obtained rules from the specified data variables. Conclusions and decision making that is reasoned based on fact and analyzing done. In data analyzing, we use the help of a smartphone. It is used to collect the necessary data, because one of the specified parameters is the strength of 4G network, chat application and battery life on smartphone. Stages in analyzing initial data among others are:

1. Data Selection
   The data are taken with some variations of certain events. Data are obtained based on survey results that conducted in the field by using smartphone. The data are gathered by monitoring the activities of the smartphone with reference variables and events that have been predetermined. Each event is limited within 15 minutes.
2. Pre-Processing
   In this process, data grouping is done into training data and testing data based on the obtained data from the survey results. The data that has absolute value and the output results that have been known are grouped into training data, while the data that have the value but the output is not known then it will be grouped into the testing data [8], [9].
3. Transformation
   In the transformation stage, the data is processed into a table format into Microsoft Excel so that the data can be read more concisely and structured. The data table or dataset will be done classification process with decision tree C4.5 algorithm.

C. Proposed Method

The proposed method is to find the rules in overcoming the problems by using classification method and C4.5 algorithm. This method is selected because it has high accuracy values and can be used appropriately because the variables have many variations and make easy for calculation by means of weighting. Steps in performing calculation of classification method using decision tree C4.5 consist of:

1. Classification
   By using classification method and decision tree of C4.5 algorithm, the data will be processed into rules by checking the weights based on the output of each data. The weights will get the Gain values in each class and the Entropy value in each attribute [5], [8], [9].
2. Decision Tree
   Based on Gain and Entropy values, the obtained values will be transformed into a tree. A tree consists of several nodes or leaf nodes that continue to take root to the bottom. The initial node is obtained from the weighted Gain of the class that has the greatest Gain value [5], [9]. To continue the calculation of the node underneath, it is done re-calculation on attributes that have values or results are not uniform. The calculations continue to be repeated until the last leaf node is found and the value of all attributes is uniform.
3. Rule
   The composition of the resulted tree then is converted into a rule. The obtained rule is come from the certainty calculation and be a calculation reference of testing data or new data that has not been known yet their output.
4. Prediction
   A prediction or evaluation is required to know the number of errors in the classification process. Error is calculated from the amount of testing data output that is different from the training data output.

D. Conceptual Framework

In this study, we have formulated a necessary conceptual framework for the basis of where the research will proceed. This stems from the problems that existed to the end of the research in the form of information that is useful for the people and extensive knowledge. The conceptual framework of this study can be divided into six discussion as in Fig. 1 below:

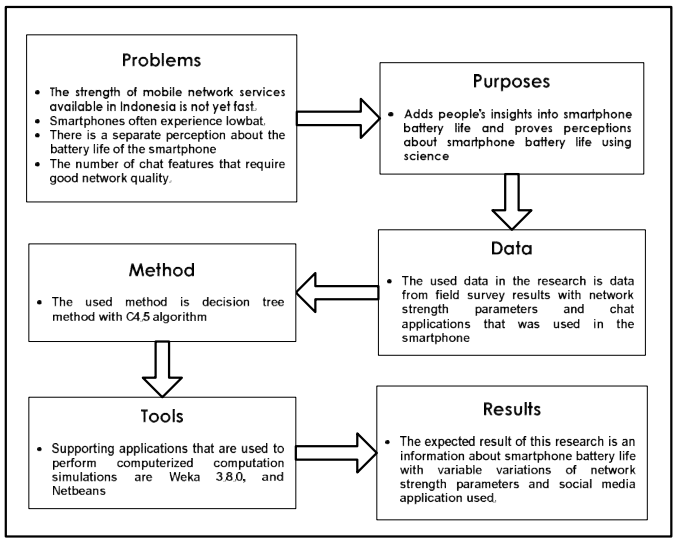


Fig. 1. Conceptual Framework.

This research is expected have benefits both for the people and science. The prospective researchers can have reference for the related research. This research is also expected to contribute in science with applied a method if data that have artificial intelligence value. The research results are expected to provide benefits to the people, especially in terms of smartphone use, so that people become wiser in using smartphones and understand a little basic mobile network.

E. Experiment and Testing Method

The conducted experiments are to simulate the calculation results that have been done by using Weka 3.8.0. This application will give several options to manipulate the output such as the number of errors, the number of nodes that are displayed and many other options that are useful for experiment. This research also conduct the tests using a developed simple application that was made by authors. The developed application was designed using the Netbeans to make coding in Java programming language [6], [10], [11]. The calculation results of simulation are used to know the comparison results of manual calculation with computerized calculation and also calculation come from developed simple program. To find out the evaluation of the classification model performance, it based on the number of dataset records that are predicted correctly and incorrectly in the classification model used containing the comparison of testing data with training data results.

III. Result and Discussion

A. Data Requirement Analysis

Analysis of data needs or data selection is done to determine the data or attributes that are used to determine the decision variables that can be shown in the TABLE I below.

TABLE I. Assessment Criteria

No.	Attributee	Criteria
1	Bar Signal	1-2 Bar 2-3 Bar 3-4 Bar
2	Weather	Sunny Cloudy Rainy
3	Location	Indoor Outdoor
4	User Type	Passive Active
5	Number of Applications	1 2 3
6	Stability	Yes No

The conducted surveys are to obtain data of the battery life attributes that has wasteful and saving criterias. It is based on the variation of predefined data of the variable criteria. This survey is done by utilizing smartphone using variable variations of data and done within 15 minutes. The survey results can be shown in the TABLE II below.

TABLE II. Survey Results

No	Signal Bar	Weather	Loc.	User Type	Apps	Stable	Batt. Life
1	1-2	sunny	outdoor	Passive	1	no	4%
2	1-2	cloudy	outdoor	Passive	1	no	5%
3	1-2	rainy	outdoor	Passive	1	no	4%
4	1-2	sunny	indoor	Passive	1	no	6%
5	1-2	cloudy	indoor	Passive	1	no	8%
6	1-2	rainy	indoor	Passive	1	no	4%
…	**…**	…	…	…	…	…	…
49	1-2	sunny	outdoor	Passive	2	yes	5%
50	1-2	cloudy	outdoor	Passive	2	yes	5%
51	1-2	rainy	outdoor	Passive	2	yes	4%
…	**…**	…	…	…	…	…	…
67	1-2	sunny	outdoor	Active	3	yes	4%
68	1-2	cloudy	outdoor	Active	3	yes	7%
69	1-2	rainy	outdoor	Active	3	yes	7%
73	2-3	sunny	outdoor	Active	1	no	3%
74	2-3	cloudy	outdoor	Active	1	no	4%
75	2-3	rainy	outdoor	Active	1	no	3%
…	**…**	…	…	…	…	…	…
94	2-3	sunny	indoor	Active	2	no	5%
95	2-3	cloudy	indoor	Active	2	no	6%
96	2-3	rainy	indoor	Active	2	no	4%
…	**…**	…	…	…	…	…	…
139	2-3	sunny	outdoor	Active	3	yes	3%
140	2-3	cloudy	outdoor	Active	3	yes	6%
141	2-3	rainy	outdoor	Active	3	yes	6%
…	**…**	…	…	…	…	…	…
145	3-4	sunny	outdoor	Passive	1	no	2%
146	3-4	cloudy	outdoor	Passive	1	no	3%
147	3-4	rainy	outdoor	Passive	1	no	2%
…	**…**	…	…	…	…	…	…
187	3-4	sunny	outdoor	Passive	1	yes	2%
188	3-4	cloudy	outdoor	Passive	1	yes	3%
189	3-4	rainy	outdoor	Passive	1	yes	4%
…	**…**	…	…	…	…	…	…
214	3-4	sunny	indoor	Passive	3	yes	1%
215	3-4	cloudy	indoor	Passive	3	yes	3%
216	3-4	rainy	indoor	Passive	3	yes	2%

Based on the survey results, calculation average is counted accordance with procentage of reduced battery life. We do not show all the data but we make conclusion that can be shown in TABLE III below.

TABLE III. CALCULATION OF BATTERY LIFE LIMIT

	Signal Bar 1-2	Signal Bar 2-3	Signal Bar 3-4	Total
Summary	400%	315%	211%	926%
Average				4,29

Based on the calculation results obtained value 4.29 which becomes the size limit that is < 4.29 value is classified into Battery Life Saving and value > 4.29 is classified into Battery Life Wasteful.

B. Classification Method Using Decision Tree

To change the obtained data into an information in the form rules, it is performed calculation using decision tree C4.5 algorithm. The result of node calculation is represented by the calculation table of the initial node where in the initial node calculation table obtained values that become the answer to the problem that is the impact of network strength and chat application to battery life of the smartphone.

The C4.5 algorithm calculation steps are generally: (a) defining the attribute as its root, (b) making a branch on each value, (c) divide case into branch and (d) repeat each branch until all the cases on the branch have the same class. Here is a formula for determining Entropy [5], [10], [12], [13], [14]:

$$\text{Entropy}\ (S) = \sum_{i=1}^{k} -\text{pi} * \log_2 pi \qquad (1)$$

Where:

S is a special set
k is the number of partitions on S
pi is a probability derived from Sum (Yes) or Sum (No) divided by a special total.

After doing the calculation of entropy, then the next step is to do the gain calculation. Here is the gain calculation:

$$Gain\ (A) = Entropi\ (S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} x\ Entropi(S_i) \qquad (2)$$

Where:

|Si| is number of cases on partition to i
|S| is number of cases on S

Based on the the calculations result, we obtained the best attribute for the initial node that is the signal bar attribute with the highest gain value 0.313294. The next step is to calculate entropy and gain returns by assigning the selected attribute criteria such as the calculation of the variation variables on the criteria present in the attributes of signals bar 1-2, 2-3, and 3-4. In the preparation of the root of the decision tree, it needs to be considered is the value of the criteria. If the value of a kind then do not need to be calculated on the value. While if the value in the criteria is still varied then need to do the calculation again. The whole processes do not appear in this paper, this is caused to many repetitions for signal bar 2-3 and 3-4. Here is for example, the initial node calculation table can be seen on TABLE IV below.

TABLE IV. INITIAL NODE CALCULATION

Node	Atributte	Criteria	Case	Saving	Wasteful	Entropy	Gain
1	Total		216	128	88	0.975119	
	Bar						0.313294
		1-2	72	18	54	0.811278	
		2-3	72	40	32	0.991076	
		3-4	72	70	2	0.183122	
	Weather						0.08305
		sunny	72	51	21	0.870864	
		cloudy	72	30	42	0.979869	
		rainy	72	47	45	0.825473	
	Location						0.002307
		Indoor	108	61	47	0.987844	
		Outdoor	108	67	41	0.95778	
	User Type						0.031255
		passive	108	75	33	0.887976	
		active	108	53	55	0.999753	
	No of Apps						0.01312
		1	72	47	25	0.931563	
		2	72	45	27	0.954434	
		3	72	36	36	1	
	Stable						0.000256
		yes	108	65	43	0.969857	
		no	108	63	45	0.979869	

From the initial node calculation table, the highest value of Gain is found in the signal bar. These results prove that the main factor that affects the battery life of the smartphone is the number of signal bars, followed by the weather, the type of application users chat and so on. In the preparation of the root of the decision tree that need to be considered is the value of the criteria, if the value of a kind then do not need to be calculated on the value. While if the value in the criteria is still varied then need to do the calculation again. After the decision tree is formed then it will be followed by the formation of rules that become the reference in determining the prediction results. The predicted results compared with the existing data are then evaluated using confusion matrix to determine the accuracy, precision and recall.

C. Designing Decision Tree

After getting the node from the weighting result, then the decision tree is made by taking the adjusting weighting condition, starting from top to bottom (root). Fig. 2 is the forming result of a decision tree with the initial node attribute signal bar 1-2 for representation of the formed decision tree in this paper. There are 51 nodes under the signal bar 1-2 which consists of attribute variations that have been adapted to the calculation conditions. Then, there are 57 node under the singnal bar 2-3 and 14 node under the signal bar 3-4. We do not show the decision tree under signal bar 2-3 and 3-4 in this paper. After the decision tree is formed, then the next step is to determine the rules that are used as a science that can be done for further implementation.

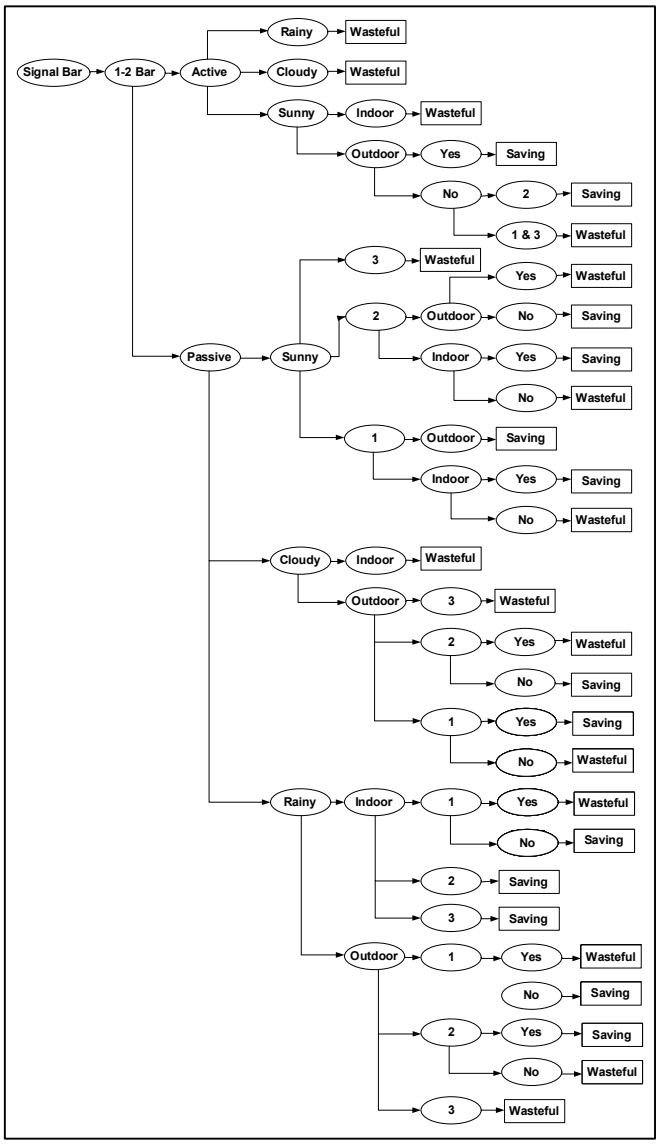


Fig. 2. Decision Tree for signal bar 1-2.

After the decision tree is formed then the next step is to determine the rules that are used as a science that can be done for further implementation. Several of the formed rules are as follow:

1. *R1= If Signal Bar = 1-2 ^ User Type = Active ^ Weather = Sunny ^ Location = Indoor then Battery Life = Wasteful*
2. *R2= If Bar Signal = 1-2 ^ User Type = Active ^ Weather = Sunny ^ Location = Outdoor ^ Stable = Yes then Battery Life = Save*
3. *R3= If Signal Bar = 1-2 ^ User Type = Active ^ Weather = Sunny ^ Location = Outdoor ^ Stable = No ^ Number of Applications = 2 then Battery Life = Save* and so forth.

There are 65 rules of the formed decision tree based on observations in the field.

D. Simulation Using WEKA

Experiments on WEKA tools are required to know and measure the strength of the used methods.

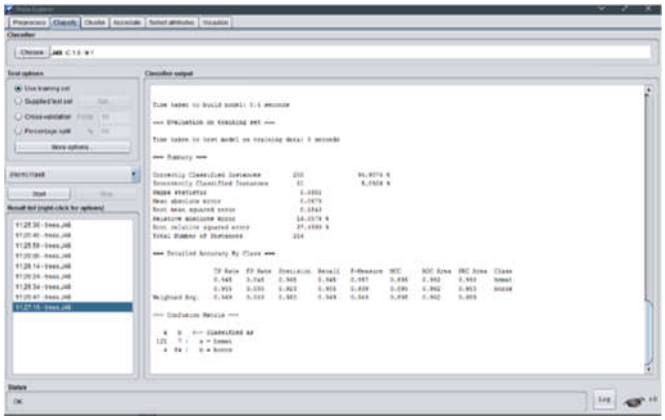


Fig. 3. Simulation the method using WEKA.

Fig. 3 is the result of the obtained calculations by using WEKA tools. The formed decision tree can be seen in the tools along with the accuracy, error and confusion matrix values.

E. Experiments Using a Developed Simple Application

Experiments are also performed using a developed simple application to calculate whether the application is eligible to be used by applying the rules of the decision tree well as shown in Fig. 4 below.

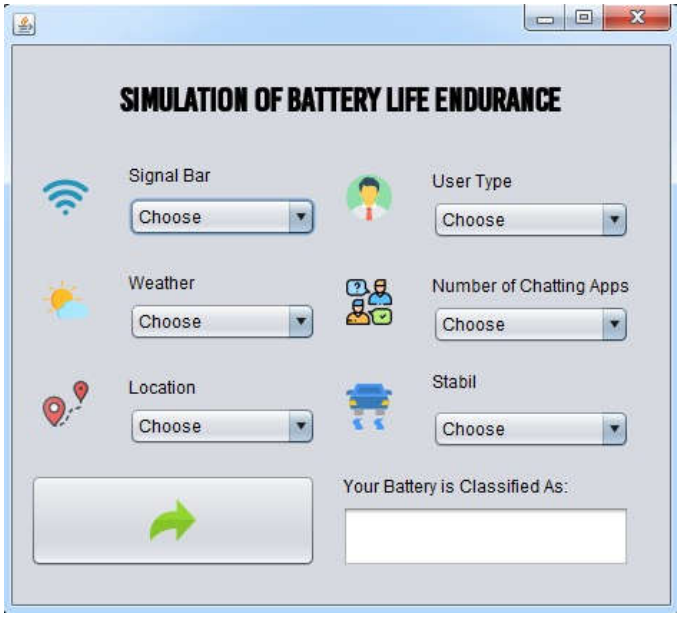


Fig. 4. Experiments using a simple application.

In Fig. 4, there are several options for assigning attribute values to each class. Attributes are filled with data testing. After all attributes in the class are filled, then the application will process with the rules that have been obtained and then produce the output of a wasteful or saving label for battery life of the smartphone.

F. Performance Evaluation

The performance evaluation of the classification model is based on the count of dataset records that are predicted correctly and incorrect using confusion matrix. It uses a table that contains the comparison results of the testing data with training data [12].

TABLE V. CONFUSION MATRIX

		Prediction	
		Class = 1	*Class* = 0
Actual	*Class* = 1	Z11	Z10
	Class = 0	Z01	Z00

Where:
Z11: Prediction is positive and actual data is positive
Z10: Prediction is negative and actual data is positive
Z01: Prediction is positive and actual data is negative
Z00: Prediction is negative and actual data is negative

$$Accuracy = \frac{Z11 + Z00}{Z11 + Z10 + Z01 + Z00} \quad (3)$$

While precision is a measure of the accuracy of a certain class that has been predicted.

$$Precision = \frac{Z11}{Z01 + Z11} \quad (4)$$

Recall is a percentage of data with a positive value whose predictive value is also positive.

$$Recall = \frac{Z11}{Z10 + Z11} \quad (5)$$

Based on the calculation done, the comparison table obtained which contains the accuracy, precision and recall values of each calculation as shown in TABLE V.

TABLE VI. PERFORMANNCE EVALUATION RESULTS

	Accuracy	**Precision**	**Recall**
Manually	100%	100%	100%
WEKA	94.91%	96.80%	94.53%
Simulation	100%	100%	100%

Based on TABLE VI, the result from manual and using simple application calculation get accuracy, precision and recall value 100%. We use manual calculation, WEKA and a developed simple application are to test the algorithm and to know the differences among them. Maximum results can be obtained because the calculation is almost completely done by researchers. While the results on the WEKA tools get smaller results because the WEKA when processing calculations to determine the node, the same values of the gain or entropy only left alone and given a value that is still mixed (not uniform) therefore the decision tree results on WEKA experience error.

IV. CONCLUSION AND FUTURE WORK

Based on the data analysis results using classification method with decision tree C4.5 that is applied to WEKA and developed application, it can be concluded as follows:

1. By applying the classification method using decision tree C4.5 to the research problem, after calculating the value of gain, entropy, forming a decision tree and determining the rules, the battery life can be classified become saving or wasteful easily.
2. In the calculation of the initial node, it can be seen that the main cause of smartphone battery life is the signal bar strength of the operator in the smartphone that has a gain of 0.313. Followed by several parameters in it that adjusted to the parameters that have been determined at the beginning node, where the weather 0.831, the location 0.023, the user type 0.313, the number of applications 0.131 and stable 0.0003.
3. This study get a comparative table data that contains the results of manual calculations, WEKA and developed simple application for experiment. Based on the comparison results among the calculations, we can conclude that the manual calculation and simple simulation applications have accuracy, precision and recall value of 100% while WEKA have accuracy value of 94.91%, precision 96.80% and recall 94.53%. (3)

In the research, we realize there are still some shortcomings that need to be done development and improvement such as: (1) objects of observation can be developed on some types of smartphones and (2) further research is needed to further detail what parameters affect the battery life of the smartphone.

REFERENCES

[1] F. Fauzi et. al., "Analisis Penerapan Teknologi Jaringan LTE 4G Di Indonesia", Majalah Ilmiah UNIKOM Vol.10, No. 2 , pp. 281-290, 2012.

[2] B. Wahyu et al., "Teknologi 4G Pada Jaringan GSM Untuk Kebutuhan Mobile Internet Di Kota Yogyakarta", Jurnal JARKOM Vol. 2 No. 2, pp. 32-43, 2015.

[3] R. K. Singh and R. Singh, "4G LTE Cellular Technology: Network Architecture and Mobile Standards", International Journal of Emerging Research in Management &Technology, Vol. 5, Issue-12, 2016.

[4] L. Budiono, "Persepsi dan Harapan Pengguna terhadap Kualitas Layanan Data pada Smartphone di Jakarta User," Buletin Pos dan Telekomunikasi, vol. 11 No.2, pp. 93-108, 2013.

[5] N. Z. Rahma and A. Setyono, "Penerapan Algoritma C4.5 Dalam Memprediksi Kesiapan Siswa SMP IT PAPB Semarang Menghadapi Ujian Nasional", Jurnal Sisfotenika Vol. 8, No. 1, pp. 35-46, 2018.

[6] T. C. Sharma and M. Jain, " WEKA Approach for Comparative Study of Classification Algorithm", International Journal of Advanced Research in Computer and Communication Engineering", vol. 2 No.4, pp. 1925- 1931, 2013.

[7] H. Umar, "Analisis Pengaruh Sinyal 3G Pada Smartphone Yang Menyebabkan Baterai Cepat Drop/Lowbat", Jurnal Inspiration, vol. 5 No.1, pp. 56-64, 2015.

[8] B. Hssina, et. al., "A comparative study of decision tree ID3 and C4.5", International Journal of Advanced Computer Science and Applications, Special Issue on Advances in Vehicular Ad Hoc Networking and Applications, 2013.

[9] W. Dai and W. Ji, "A MapReduce Implementation of C4.5 Decision Tree Algorithm", International Journal of Database Theory and Application, Vol. 7, No. 1, 2014.

[10] Susanti, "Klasifikasi Kredit Menggunakan Metode Decision Tree Pada Nasabah Pada BPR BKK Gabus", 2014.

[11] S. Afrizal, "Rancang Bangun Aplikasi Dekstop Kamus Indonesia, Inggris Dan Arab Menggunakan Netbeans Dan Mysql", Jurnal Teknilk Informatika Politeknik Sekayu (TIPS) , Vols. Volume 1, No. 1, pp. 1-9, 2014.

[12] B. Santoso, "Analisis Dan Penerapan Metode C4.5 Untuk Prediksi Loyalitas Pelanggan", Jurnal Ilmiah Fakultas Teknik LIMITS, Vol. 10 No 1, 2014.

[13] D. T. Larose, "Discovering Knowledge in Data : An Introduction to Data Mining", John Wiley & Sons, 2005.

[14] G. L. Agrawal1 and H. Gupta, "Optimization of C4.5 Decision Tree Algorithm for Data Mining Application", International Journal of Emerging Technology and Advanced Engineering, Vol. 3, No. 3, 2013

The Performance of FOSMC and Boundary – SMC in Speed Controller and Current Regulator for IFOC-Based Induction Motor Drive

Angga Wahyu Aditya[1], Muhammad Rizani Rusli[2], Bayu Praharsena[3], Era Purwanto[4], Dedid Cahya Happyanto[5], and Bambang Sumantri[6]
[1,2,3,4,5,6]Postgraduate of Electrical Engineering
[4]Industrial Electrical Engineering
[5,6]Electronics Engineering
Electrical Engineering Department, Politeknik Elektronika Negeri Surabaya, Surabaya, Indonesia
[1]wahyua800@gmail.com, [2]ruslirizani@gmail.com, [3]bayupraharsena16@gmail.com, [4]era@pens.ac.id, [5]dedid@pens.ac.id, [6]bambang@pens.ac.id

***Abstract* – Indirect Field-Oriented Control (IFOC) is the part of vector control method which is suitable for industrial application especially for induction motor (IM) drives. It causes IFOC doesn't need flux sensor which changes the construction of IM like Direct Field-Oriented Control (DFOC). Unfortunately, the robustness and the stability of the system cannot be guaranteed by IFOC. Sliding Mode Control (SMC) is a controller which can guarantee both of them. Ordinary SMC or First-Order SMC (FOSMC) has disadvantages in chattering phenomenon because of discontinuous control input (sign(.) function). This chattering phenomenon also consumes more power to a standstill the robustness. Chattering phenomenon can be reduced by giving thin boundary layer (Boundary – SMC). Boundary – SMC using saturation function proposed to reduced chattering phenomenon and power consumption. This paper discusses FOSMC and Boundary – SMC applied for speed controller (SC) and direct-current regulator in 10HP IM. The results show power consumption using Boundary – SMC is lower than FOSMC (real power decreases 21.26%, active power decreases 44.27% and apparent power decreases 25.49%). Power factor analyze shows FOSMC in direct-current regulator is the best but consume more real power than the others.**

Index Terms - First Order SMC, Boundary–SMC, IFOC, IM, Power Evaluation.

I. INTRODUCTION

The IM has been used in the industrial world for many years due to its reliability and low price [1], [2]. There are two methods for IM drives: scalar control and vector control. In one hand, scalar control is easy to use and application. In another hand, it has low efficiency. Vector control method widely used in order to increase the efficiency of the IM. IFOC is one vector control method is suitable for industrial application [3].

IFOC which is part of vector control is a suitable system to control IM with high-performance AC drives [4]. In IFOC system, a complicated mathematical model of the IM can be represented in a similar manner to DC motor [3].

SMC uses as the robust controller can guarantee robustness and stability of the system. In SMC, the stability of the system guaranteed by Lyapunov stability theory and the robustness by sign(.) function. However, FOSMC has disadvantages in chattering phenomenon. This chattering phenomenon consumes more power to a standstill the robustness. Thin boundary layer designed in chattering area to reduced the chattering phenomenon (Boundary – SMC) using saturation function [5]. This technique also decreased power consumption [6].

In this work, PI controller compared with FOSMC used in direct-current regulator; direct-current regulator and SC; and Boundary – SMC used in SC; Boundary – SMC used in SC while direct-current regulator using FOSMC in order to analyze the performance (Root Mean Square (RMS) direct stator response, rotor speed response and power factor) and power consumption (real power, active power, and apparent power) of IFOC-based 10HP IM drives. Boundary – SMC proposed to reduced power consumption and increase power factor.

II. IM MODEL

The electrical model of IM in d-q rotating frame is [3] :

$$\frac{di_{sd}}{dt} = a_1 i_{sd} + \omega_s i_{sq} + a_2 \varphi_{rd} + a_3 \omega_r \varphi_{rq} + a_4 U_{sd}$$

with $a_1 = \frac{R_s L_r^2 + R_r L_m^2}{L_r w}$; $a_2 = \frac{R_s L_m}{L_r w}$; $a_3 = \frac{L_m}{w}$; $a_4 = \frac{L_r}{w}$;

$w = \sigma L_r L_s$; $\sigma = 1 - \frac{L_m^2}{(L_s + L_r)}$ is total leakage factor.

The equation of speed of IM is :

$$\frac{d\omega_r}{dt} = \frac{L_m}{J L_r}(\varphi_{rd} i_{sq} - \varphi_{rq} i_{sd}) - \frac{1}{J} m_0$$

Where i_s and U_s are current and voltage of stator vectors in d-q rotating frame, φ_r is rotor flux vectors in d-q rotating frame, m_0 is torque load and ω_r is rotor speed.

III. IFOC-BASED IM DRIVES

Fig. 1 shows the basic concept of IFOC for an IM. This system makes the electromagnetic torque (T_e) and rotor flux (φ_r) can control separately by direct-axis stator current (i_{sd}) and quadrature-axis stator current (i_{sq}) shown in Fig. 2.

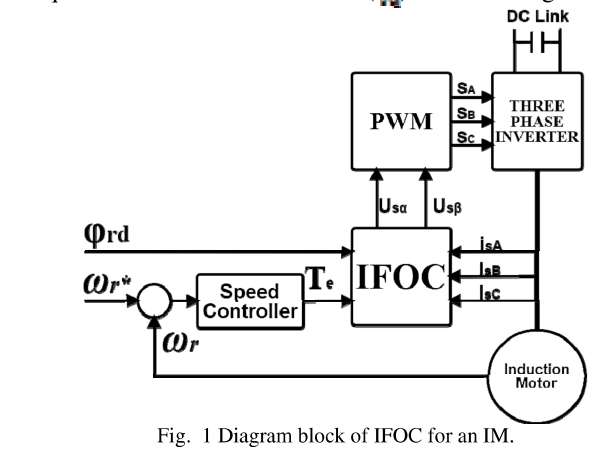


Fig. 1 Diagram block of IFOC for an IM.

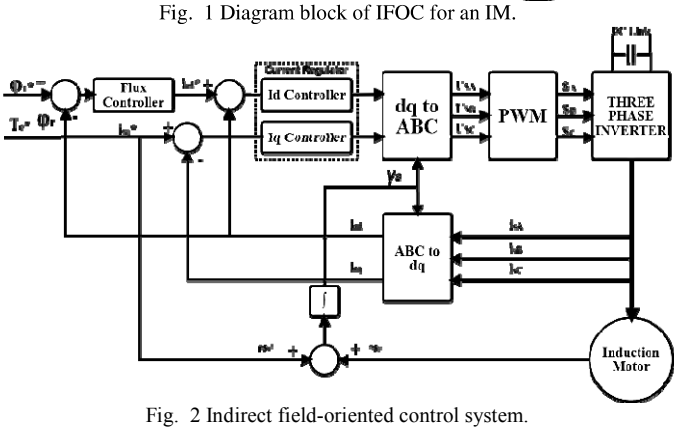


Fig. 2 Indirect field-oriented control system.

By making the rotor flux vectors of electrical and equation of speed IM model to be constant, the equation of IFOC is [3], [7], [8] :

$$\frac{di_{sd}}{dt} = a_1 i_{sd} + \omega_s i_{sq} + a_2 \varphi_{rd} + a_4 U_{sd}$$

The equation of speed become :

$$\frac{d\omega_r}{dt} = \frac{L_m}{J L_r} \varphi_{rd} i_{sq} - \frac{1}{J} m_0$$

The electromagnetic torque equation is :

$$T_e = \frac{3}{2} P \frac{L_m}{L_r} \varphi_{rd} i_{sq}$$

The correlation between equation of speed and electromagnetic torque in IFOC is :

$$\frac{d\omega_r}{dt} = \frac{2}{3 J} P T_e - \frac{1}{J} m_0$$

The flux vector position angle is [7] :

$$\theta = \int_0^t (P\omega_r + \omega_{sl})\, dt$$

IV. SLIDING MODE CONTROL (SMC)

The sliding surface to design SMC which presented by [9], [10] is :

$$S(e;t) = \left(\frac{d}{dt} + \lambda\right)^{n-1} e \qquad (1)$$

where n is a sliding surface degree, $e = X_{desire} - x$ is an error and λ is constant.

The FOSMC controller design follows:

$$U_{smc} = U_{eq} + U_n$$

where U_{smc} is the SMC controller, U_{eq} is the equivalent controller and U_n is added to guarantee the attractiveness of sliding surface in (1).

The Lyapunov function to guarantee the stability of system follows :

$$V = \frac{1}{2} S^2 \qquad (2)$$

The derivative of (2) is :

$$\dot{V} = S\dot{S}$$

The equivalent control signal (U_{eq}) of FOSMC designed using Lyapunov stability theory where the derivative of a Lyapunov function is zero that shown in (3).

$$U_{eq} = \dot{V} = 0$$

or

$$0 = S\dot{S} \qquad (3)$$

From the current and speed equation of IFOC, (1) and (3), the FOSMC controller of direct-current regulator and Speed are :

The FOSMC design of direct-current regulator is :

$$U_{eq}(U_{sd}) = \frac{1}{a_4}\left[\dot{i}_{sd}^{*} + a_1 i_{sd} - \omega_s i_{sq} - a_2 \varphi_{rd}\right]$$

$$U_n = K\left[S_{i_{sd}}(e,t)\right] + \beta\, \mathrm{sign}\left[S_{i_{sd}}(e,t)\right]$$

The FOSMC design of SC is :

$$U_{eq}(T_e) = \frac{3 J}{2}\left[\dot{\omega}_r^{*} + \frac{1}{J} m_0\right]$$

$$U_n = K\left[S_{\omega_r}(e,t)\right] + \beta\ \mathrm{sign}\left[S_{\omega_r}(e,t)\right]$$

where $S_x(e_x,t) = \lambda e_x$, K and β is a positive constant gain.

Boundary technique aims to reduce the chattering phenomenon in FOSMC cause of discontinuous input. This chattering consumes more power to hold the robustness. Thin boundary layer restricts the discontinuous input. But, It cannot guarantee the robustness and stability of the system in the boundary area. The boundary equation shows in (4) [6], [11].

$$\mathrm{sat}(S) = \begin{cases} \mathrm{sign}(S) & \text{for } |S| > \beta_1 \\ \dfrac{S}{\beta_1} & \text{for } |S| \le \beta_1 \end{cases} \quad (4)$$

where β_1 is boundary layer constant.

By substitute (4) to the SC in FOSMC design of IFOC system, the Boundary – SMC of the SC is :

$$U_{eq}(T_e) = \frac{3}{2}J\left[\dot{\omega}_r^* + \frac{1}{J}m_0\right]$$

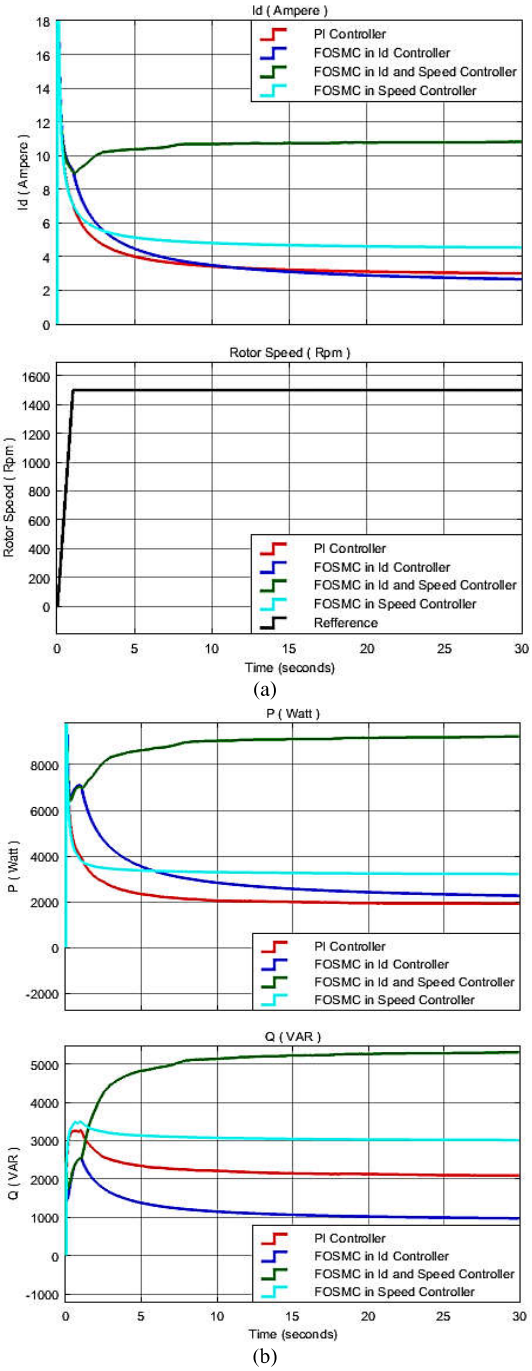
$$U_n = K\left[S_{\omega_r}(e,t)\right] + \beta\ \mathrm{sat}\left[S_{\omega_r}(e,t)\right]$$

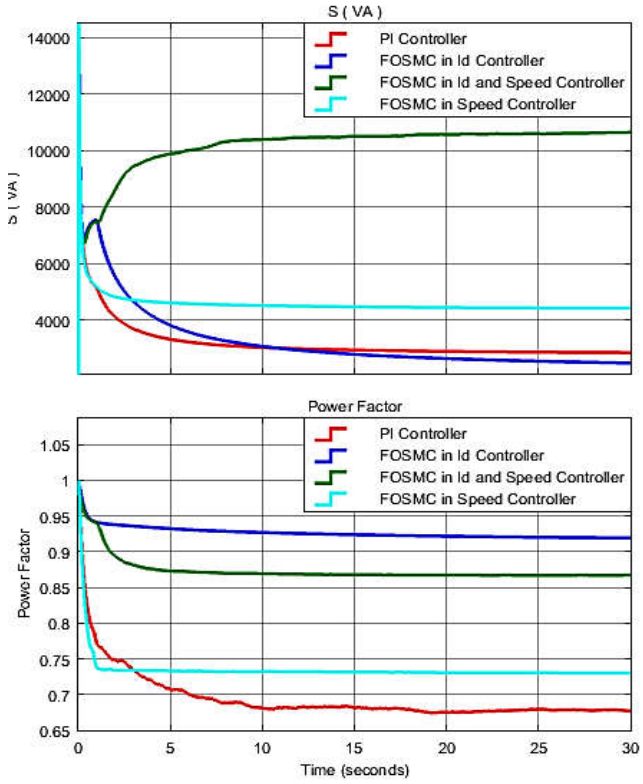
V. RESULT AND DISCUSSION

This section presents the results of the studied system. To analyze the performance of FOSMC and Boundary – SMC in IFOC can be reviewed from direct-stator current, speed response, power consumption (real, reactive and apparent power) and power factor. The whole system is simulated using MATLAB/Simulink with 10HP IM under constant speed reference signal and variable torque load (10 Nm at 10 seconds). However, the torque load in the first simulation is 0 Nm (no load condition).

$$P = \frac{3}{2}\left[\left(U_{sd}i_{sd}\right) + \left(U_{sq}i_{sq}\right)\right] \quad (5)$$

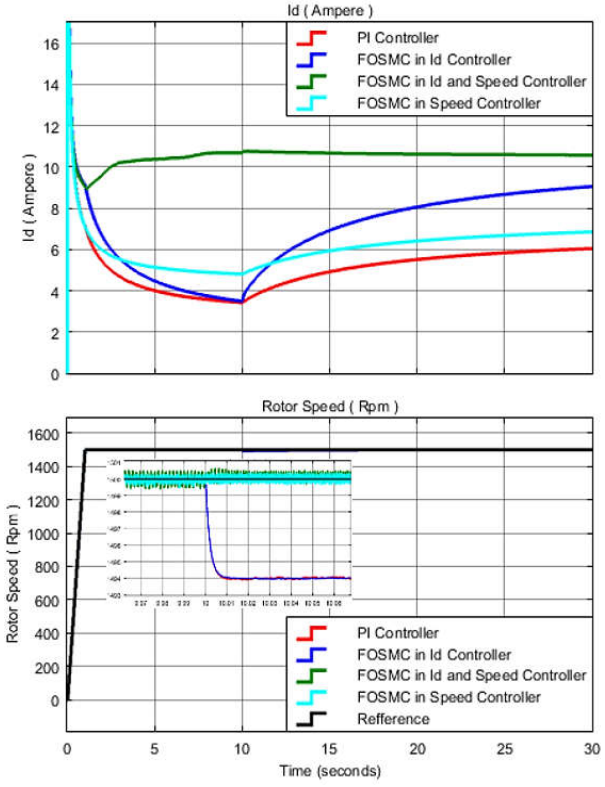
Fig. 3 and Fig. 4 show the performance of FOSMC designed in Direct-current regulator; SC; Id and SC compared with PI controller with 0 Nm and 10 Nm of torque load. In one hand direct-stator current the FOSMC designed in Id and SC shows the best performance with 2.94 % of impairment in torque load = 10 Nm. in another hand, it has the biggest RMS value of direct-stator current. The lowest RMS value is FOSMC designed in Direct-current regulator with 2.62 A. FOSMC designed in SC; Id and SC has better performance in rotor speed response with no changes of values. Power consumption calculated using (5) [12], [13], [14] and power triangle uses to calculate reactive and apparent power. FOSMC designed in Id and SC consume more power (real, reactive and apparent power) than the others. But the value is decreasing (real power decrease 12.16 %, active power decrease 18.63 and apparent power decrease 14.86 %) in under disturbance (torque load = 10 Nm) while the others are increasing. Power consumption under 10 Nm of torque load using Boundary – SMC in SC and FOSMC in Direct-current regulator is lower than FOSMC in speed and Direct-current regulator (real power decrease 21.26%, active power decrease 44.27% and apparent power decrease 25.49%). In power factor side, FOSMC designed in Direct-current regulator has the best performance in power factor with 0.92 in a no-load condition and 0.91 in 10 Nm of torque load. Unfortunately, it consumes more power than Boundary – SMC in SC and FOSMC in Direct-current regulator which shown in Fig. 5 and 6. Another disadvantage is the robustness in speed response cannot guarantee which shown in Fig. 6 (a).



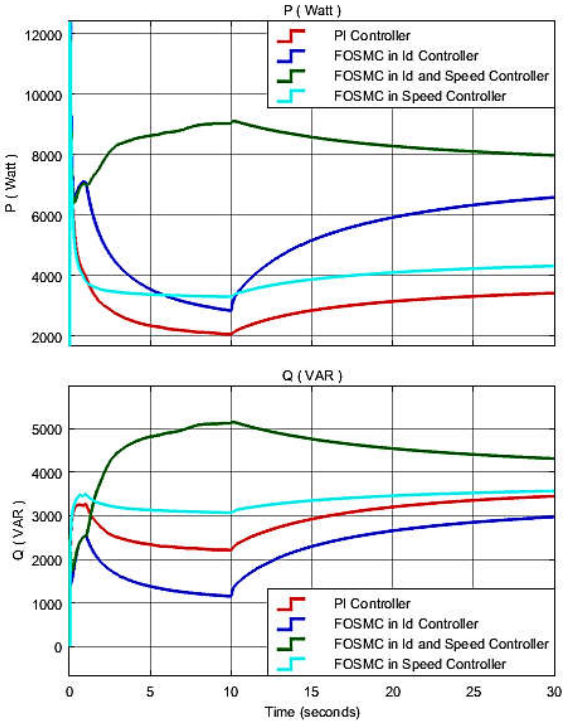


(c)

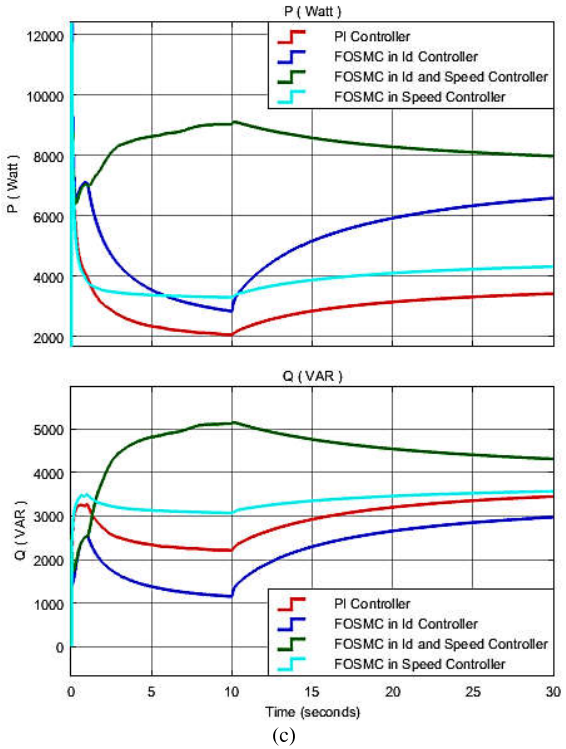
Fig. 3 Comparison between (1) PI , (2) FOSMC in Direct-current regulator, (3) FOSMC in SC and (4) FOSMC in Id and SC with torque load = 0 Nm. (a) RMS Id (A) and rotor speed (Rpm) (b) RMS real power (Watt) and RMS reactive power (VAR) (c) RMS apparent power (VA) and power factor.



(a)



(b)

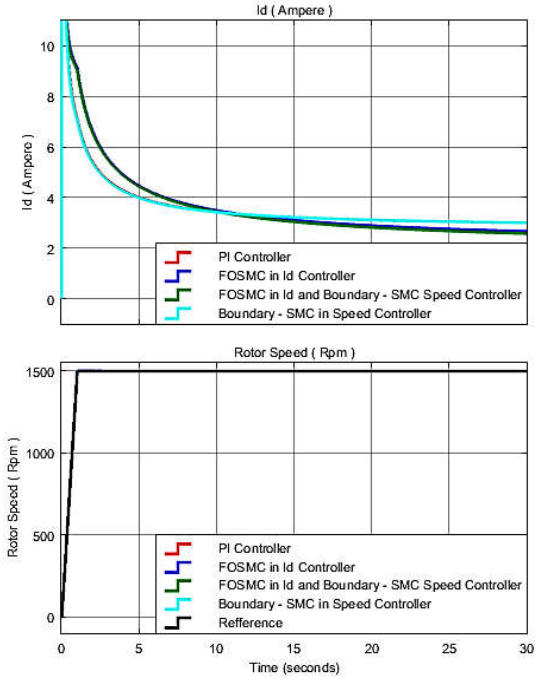


(c)

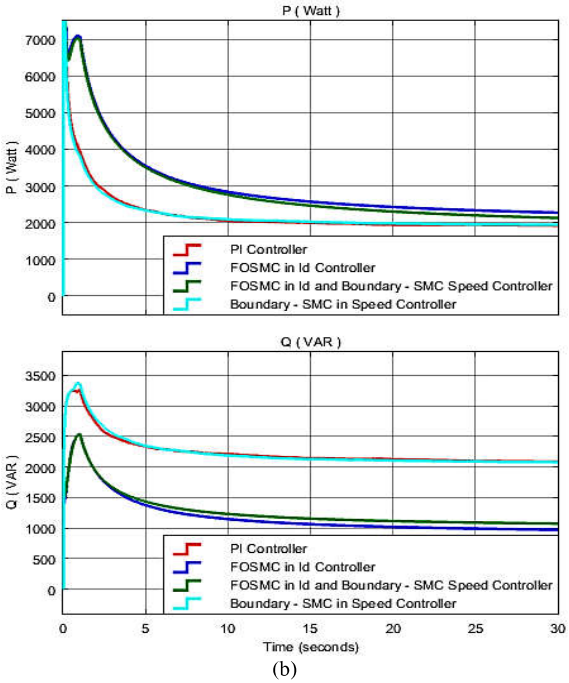
Fig. 4. Comparison between (1) PI, (2) FOSMC in Direct-current regulator, (3) FOSMC in SC and (4) FOSMC in Id and SC with torque load = 10 Nm. (a) RMS Id (A) and rotor speed (Rpm) (b) RMS real power (Watt) and RMS reactive power (VAR) (c) RMS apparent power (VA) and power factor.

Fig. 5 and Fig. 6 shows the performance of IFOC using: PI controller, FOSMC designed in Id, Boundary – SMC designed in SC and FOSMC in Id, and Boundary – SMC designed in the SC. Boundary – SMC gives better performance than FOSMC in Id, rotor speed, power

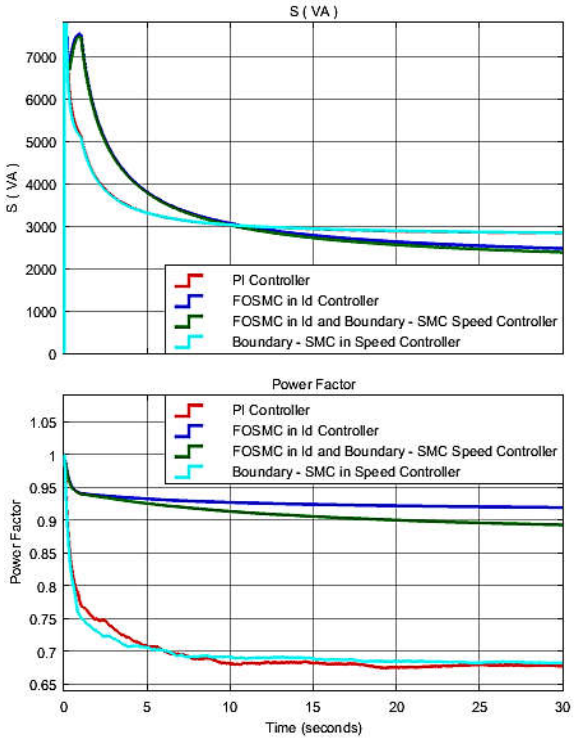
consumption and power factor. RMS value of Id using Boundary – SMC in SC and FOSMC in Direct-current regulator decrease more than three times than using FOSMC in Id and SC. In speed response, there are no significant changes by using Boundary – SMC and FOSMC designed in the SC. Boundary – SMC has decreasing value in power consumption than FOSMC. In power factor side, Boundary – SMC is better than FOSMC.



(a)

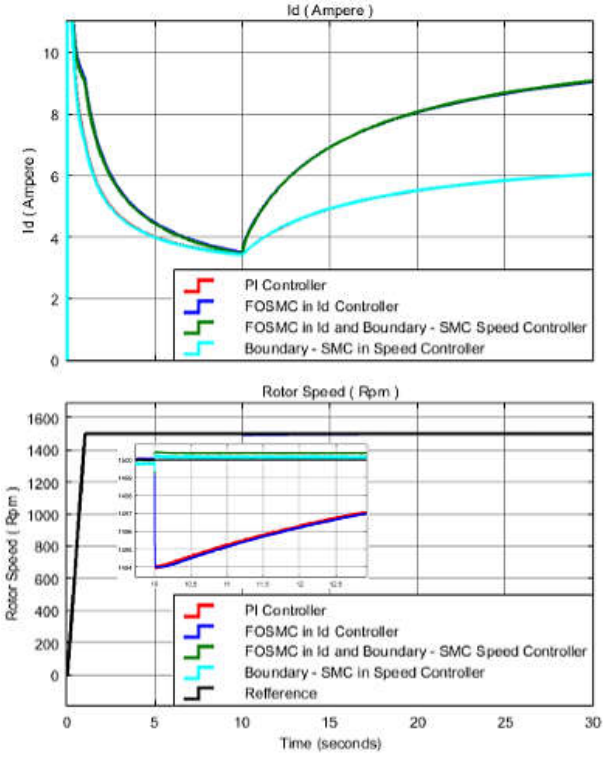


(b)



(c)

Fig. 5. Comparison between (1) PI , (2) FOSMC in Direct-current regulator, (3) Boundary-SMC in SC and FOSMC in Direct-current regulator and (4) Boundary-SMC in SC with torque load = 0 Nm. (a) RMS Id (A) and rotor speed (Rpm) (b) RMS real power (Watt) and RMS reactive power (VAR) (c) RMS apparent power (VA) and power factor.



(a)

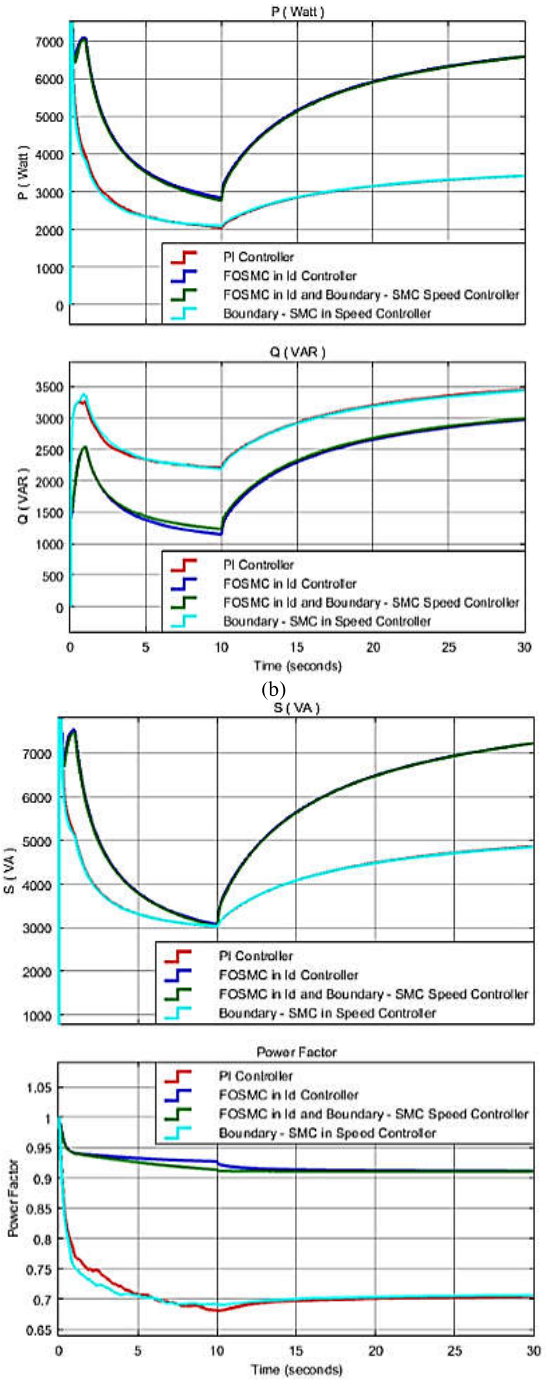


Fig. 6. Comparison between (1) PI, (2) FOSMC in Direct-current regulator, (3) Boundary-SMC in SC and FOSMC in Direct-current regulator and (4) Boundary-SMC in SC with torque load = 10 Nm. (a) RMS Id (A) and rotor speed (Rpm) (b) RMS real power (Watt) and RMS reactive power (VAR) (c) RMS apparent power (VA) and power factor.

VI. CONCLUSION

This paper discusses the development of FOSMC and Boundary – SMC in IFOC-based 10HP IM drives. FOSMC (designed in direct-current regulator, SC and Id and SC) and Boundary – SMC (SC) as robust controller compared with PI. Based on Fig. 3 – 6, Boundary – SMC designed in SC with FOSMC in Direct-current regulator is better in power consumption (real power, reactive power, and apparent power) and power factor than FOSMC in speed and Direct-current regulator. In the case of power factor analysis, FOSMC in Id has the best performance with 0.92 in a no-load condition and 0.91 when the torque load = 10 Nm. in one hand it consumes more power (real power) than the others. In another hand, the reactive power is the lowest.

ACKNOWLEDGMENT

The authors would like to thanks to parents and Electric Drive research group for support and pray. Then thanks to Electric Machine and Control laboratory – PENS.

REFERENCES

[1] A. V. R. Teja, V. Verma and C. Chakraborty, "A New Formulation of Reactive Power Based Model Reference Adaptive System for Sensorless Induction Motor Drive," *IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS,* vol. 62, no. 11, pp. 6797 - 6808, 2015.

[2] B. Praharsena, E. Purwanto, A. Jaya, M. R. Rusli, H. Toar and Ridwan, " Stator Flux Estimator Using Feed-Forward Neural Network for Evaluating Hysteresis Loss Curve in Three Phase Induction Motor," *EMITTER International Journal of Engineering Technology,* vol. 6, no. 1, pp. 168-184, 2018.

[3] H. Abu-Rub, A. Iqbal and J. Guzinski, High Performance Control of AC Drives with MATLAB/Simulink Models, West Sussex: John Wiley & Sons Ltd, 2012.

[4] B. K. Bose, Modern Power Electronics and AC Drives, Prentice Hall, 2001.

[5] A. Saghafinia, H. W. Ping and M. N. Uddin, "Sensored Field Oriented Control of a Robust Induction Motor Drive Using a Novel Boundary Layer Fuzzy Controller," *SENSORS,* vol. 13, no. 12, pp. 17025 - 17056, 2013.

[6] S. Bambang, U. Naoki and S. Shigenori, "Least square based sliding mode control for a quad-rotor helicopter and energy saving by chattering reduction," *ELSEVIER Mechanical Systems and Signal Processing,* Vol. 66-67, pp. 769-784, 2016.

[7] B. M. Wilamowski and J. D. Irwin, Second Edition : The Industrial Electronics Handbook - Power Electronics and Motor Drives, CRC Press, 2011.

[8] I. Ferdiansyah, E. Purwanto and N. A. Windarko, " Fuzzy Gain Scheduling of PID (FGS-PID) for Speed Control Three Phase Induction Motor Based on Indirect Field Oriented Control (IFOC)," *EMITTER International Journal of Engineering Technology,* vol. 4, no. 2, pp. 237-258, 2016.

[9] J.-J. E. Slotine and W. Li, Applied Nonlinear Control, New Jersey: Prentice Hall, 1991.

[10] A. W. Aditya, D. C. Happyanto and B. Sumantri, "Application of Sliding Mode Control in Indirect Field Oriented Control (IFOC) for Model Based Controller," *EMITTER International Journal of Engineering Technology,* vol. 5, no. 2, pp. 255-269, 2018.

[11] S. Lekhchine, T. Bahi, H. Bouzeria and S. Lachtar, "Robust Sliding Mode SC for Induction Motor Drive System," *International Journal Industrial Electronics and Devices,* vol. 2, no. 2, pp. 84-90, 2015.

[12] B. Kumar, Y. K. Chauhan and V. Shrivastava, "Performance Analysis of Induction Motor Drive with Optimal Rotor Flux for Energy Efficient Operation," in *International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)*, 2014.

[13] H. Sarhan, "Energy Efficient Control of Three-Phase Induction Motor," *Energy and Power Engineering,* vol. 3, pp. 107-112, 2011.

[14] H. Sarhan, "Modeling and Simulation of Optimized, PIC-Controlled, Three-Phase, Squirrel-Cage Induction Motor Drive," *International Journal of Artificial Intelligence and Mechatronics,* vol. 4, no. 1, pp. 13 - 20, 2015.

Comparison of First Order Differential Algorithm, Perturb and Observe (P&O) and Newton Raphson Methods for PV Application in DC Microgrid Isolated System

Farid Dwi Murdianto
Department of Electrical Engineering
Electronics Engineering
Polytechnic Institute of Surabaya
Surabaya, Indonesia
E-mail: farid@pens.ac.id, fariddm@ieee.org

Ainur Rofiq Nansur
Department of Electrical Engineering
Electronics Engineering
Polytechnic Institute of Surabaya
Surabaya, Indonesia
E-mail: rofiq@pens.ac.id

Yusita Heriani
Department of Electrical Engineering
Electronics Engineering
Polytechnic Institute of Surabaya
Surabaya, Indonesia
E-mail: yusitaheriani@gmail.com

***Abstract* — The necessary for energy always increases every year, so it is necessary energy alternative to overcome the phenomenon. From some alternative energy that is currently widely developed is solar energy. Unfortunately the use of solar energy with the help of solar panels produces power that varies according to the irradiation received and the temperature on the solar panel. Irradiation is the emission of energy coming from the sun. This parameter causes the output power characteristic curve to be non-linear. And the power generated by PV has not reached the maximum power point. To overcome this problem requires an arrangement to find the maximum power point and make it stable at that point. Therefore, this paper proposes "Comparison of First Order Differential Algorithms, Perturb and Observe(P&O) and Newton Raphson Methods for PV Application in DC Microgrid Isolated System" so that the maximum power point can be achieved quickly and accurately. In this paper we tried to compare 3 methods of First-order Differential, P & O (Perturb and Observe) and Newton Raphson by using ZETA Converter as its object in switching settings. By using the simulation can be obtained the results of tracking speed and maximum power level. The results of the simulation show that error power the First-order Differential is 4%, P & O is 0.26389% and Newton Raphson is 0.45178%. It is evident that Raphson's newton algorithm is superior to speed and accuracy.**

Keywords : Solar Panel, Maximum Power Point Tracking (MPPT), First-Order Differential, Perturb and Observe (P&O), Newton Raphson Method.

I. INTRODUCTION

Technological advances and industry growth make the demand for electrical energy growing every year. As we know, fossil fuels as the main source of power generation are very few. So much use of renewable energy to solve the problem. Solar energy is one of the energy that is currently being promoted its use because of its abundance. It takes solar panels to convert solar light energy into electrical energy. But the energy transfer generated by solar panels is still relatively weak. The solar cell output power is a nonlinear function against temperature and light intensity [1]. MPPT is one of the methods used to improve the effectiveness of energy harvesting using PV [2]. In Trishan's research, E (2007) described 19 MPPT algorithms including Hill Climbing, Perturb and Observe, Fuzzy Logic Control, Neural Network, ANFIS [3,4] etc. The MPPT algorithm can be used to trace the maximum power that photovoltaic can generate [5]. P&O is one of the most useful methods used. The power output from using this method is better than without using the tracking method.

II. SYSTEM DESCRIPTION

A photovoltaic cell is basically a PN semiconductor junction diode and this cell converts solar energy into electricity [6]. Generally, a PV module consists of 36 cells or 72 cells. In this system used PV with 100 Watt power as can be seen in table 1. Block diagram of the system used is shown in Fig. 1.

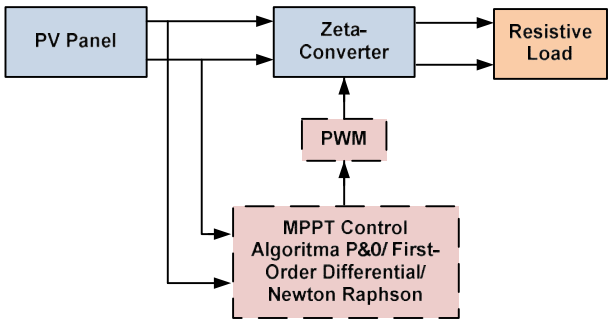


Fig. 1. Block Diagram of PV System

The output of the solar panel is a zeta converter input controlled by the algorithm to determine the exact PWM duty cycle to obtain the maximum power point.

A. Pv modelling

Photovoltaic is a tool that serves to convert or convert solar energy into electrical energy. The PV power output depends on sunlight irradiation and the temperature received by PV. The PV model is described as a simple circuit consisting of a current source representing a solar panel connected in parallel with the diode and two resistors as shown in Fig. 2. The mathematical expression of each cell are shown in equation 1 until 5.

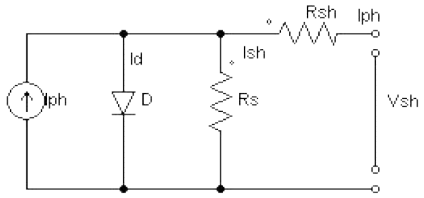


Fig. 2. The series of solar cell equations

$$I_L = I_{PH} - I_S\left(exp\left(\frac{qVd}{AN_{ST}}\right)-1\right)-\frac{(Vd)}{R_{sh}} \quad (1)$$

$$I_{PH} = \left(I_{SC} - K_i\left(T_c - T_{Ref}\right)\right)G \quad (2)$$

$$I_S = I_{RS}\left(\frac{T_c}{T_{Ref}}\right)^3 exp\left[qE_g\left(\frac{1}{T_{Ref}}-\frac{1}{T_c}\right)/K_BA\right] \quad (3)$$

$$I_{RS} = \frac{I_{SC}}{\left[exp\left(\frac{qV_{OC}}{N_SKAT_C}\right)-1\right]} \quad (4)$$

$$I_L = N_pI_{ph} - N_pI_S\left[exp\left(qV/N_SKT_CA\right) - 1\right] \quad (5)$$

Where

I_{PH}	Photo-current
I_L	Current at the output terminal
I_S	Saturation current of the diode
I_{RS}	Reverse saturation current
T_C	Cell working temperature
T_{Ref}	Reference temperature
R_{RS}	Series Resistance
K	Boltzmann's Constant (1.3806.10-23 J.K-1)
Vd	Voltage of diode
G	Solar Insolation
q	Electron charge

The above equations 1 to 5 are used in simulations using the PSIM software to obtain the characteristics of the solar cell output, as shown in Figures 3 and 4. This curve shows that the characteristics of the solar cell output are nonlinear and strongly influenced by solar radiation and temperature. In general, there is a maximum point on the V-I curve or the V-P curve, called Maximum Power Point (MPP). Where on at that point, solar cells work at maximum efficiency and produce the most output power. The corresponding values for voltage and current in each maximum power are called Vmp and Imp [7].

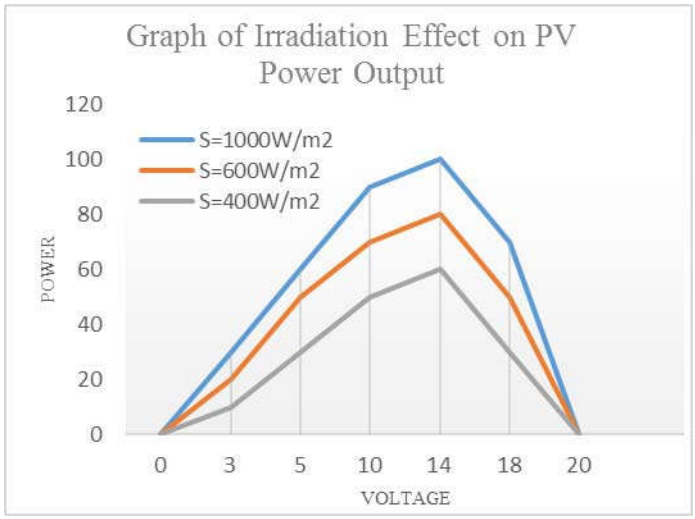


Fig. 3. Graph of Irradiation Effect on Power

Fig. 3 shows the effect of irradiation on power and output voltage of solar cell. The larger the irradiation makes the solar panel output power bigger.

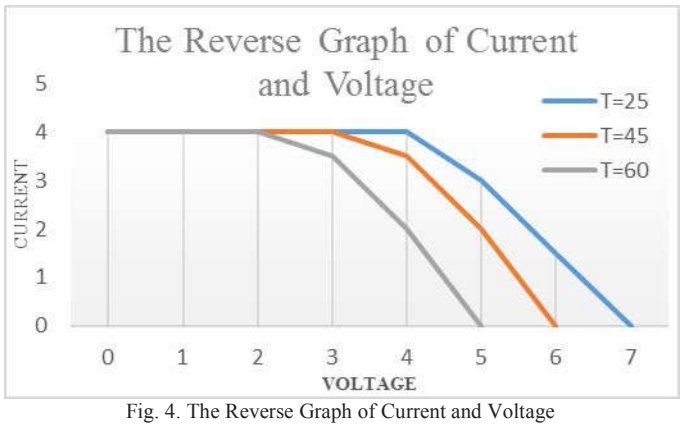


Fig. 4. The Reverse Graph of Current and Voltage

While in Fig. 4 shows the effect of the surface temperature of solar cells on curve IV. When the irradiance is kept constant at 1000 W/m^2 and temperature varies from 25ºC to 60ºC. With the increment in the temperature short circuit current increases but the open circuit voltage of cell decreases.

So the I-V characteristics shift to the left to previous curve. Power output of cell is also decreased [8].

B. *The P & O Method*

As it is known that the characteristics of solar cell output power are influenced by solar radiation and surface temperature of solar cells, an algorithm is needed to find the maximum power point (MPP) and maintain at that point of work. P & O is also called the hill climbing method [5], which refers to the V-P characteristics of the solar cells. As in Fig. 3. Flowchart of this algorithm as follows:

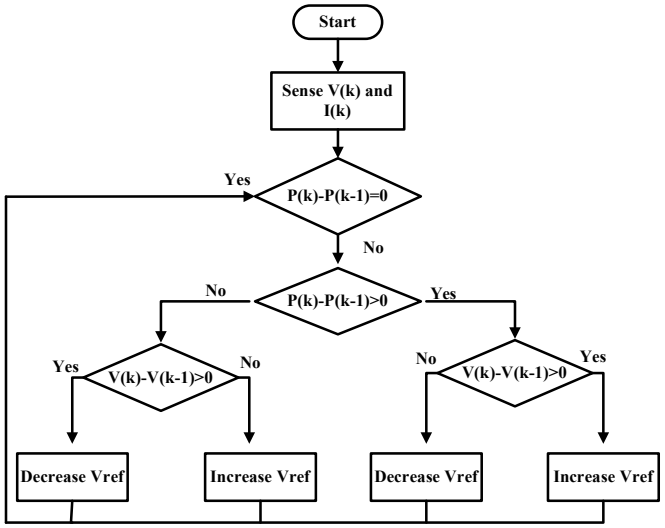


Fig. 5. Perturb and Observe Algorithm

From this flowchart the algorithm starts by measuring the voltage V (k) and current I (k) to obtain P (k). Perturb d (V) is given to observe the output power value P (k). The value of P (k) is then compared with value of P (k-1). If the value of P (k) is greater than P (k-1) it can be concluded that the perturb value performed is correct. Conversely, if the value of P (k) is smaller than P (k-1) then perturb should be done in the reverse direction. Thus the value of Maximum Power Point (MPP) can be obtained.

C. *Newton-Raphson Algorithm*

Newton method is one method to approach the equation of f(x) 0 = iteratively. The Newton Raphson method uses a starting derivative and its derivative value for subsequent gain. This method is based on the linearization of nonlinear equations and solving yields repeated linear equations [9]. In Newton Raphson's method we determine the value of almost the initial root of the so-called x0 and then determine the value of x1, x2, ...xn where the value of xn is a value almost closer to the true root value (exact solution). The value of xn is obtained by the formula

$$x_n = x_{n-1} - \left(\frac{f(x_{n-1})}{f'(x_{n-1})}\right) \quad (6)$$

The closer the initial value to the true root value the faster convergent. The calculation process can be based on absolute difference, relative error, or maximum iteration. The calculation process can be stopped when the known error is smaller than the error in the calculation or has reached the maximum iteration.

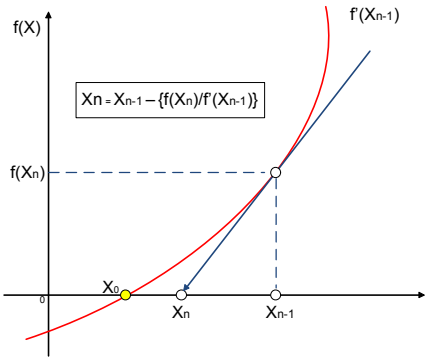


Fig. 6. Iteration Graph of Newton Raphson Method

The proposed method using a Newton Raphson variable, will allow us to compute the current Ipv with the initial value $x^{(0)}$ = I_{ph} as shown in Fig. 7.

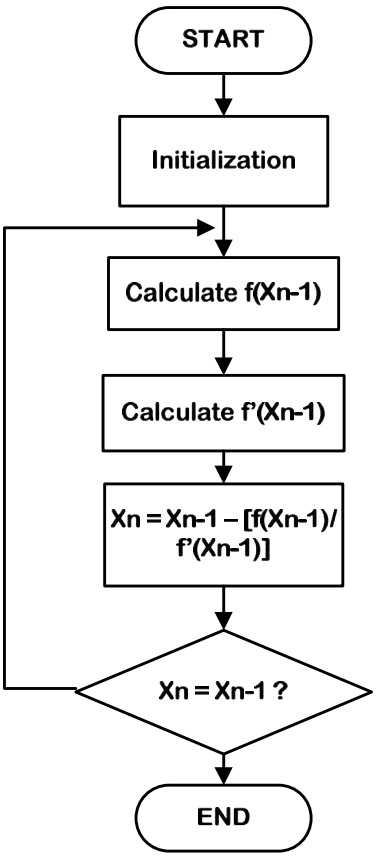


Fig. 7. Flowchart of The Proposed Method

From the flowchart in figure 7 used to get the highest point of the P-V chart. By initializing the current and the output

voltage of the solar panel will be obtained solar panel output power. Formula $x_n = x_{n-1} - \left(\frac{f(x_{n-1})}{f'(x_{n-1})}\right)$ will iterate repeatedly until it gets the highest point of output power of the solar panel.

D. ZETA CONVERTER

Fig. 8 is an image of the Zeta converter topology. Due to a single switch, this converter has very good efficiency and offers boundless region for MPPT [9][24-25]. This converter is operated in continuous conduction mode (CCM) [26] resulting in a reduced stress on its power devices and components. The Zeta converter consists of a Mosfet as a switch, a diode, two capacitors, two inductors and a load. In Zeta converter there are two state circuit in one switching period. The first is when the switch is in conduction condition (ON), and the second is when the switch is in the condition of not conduction (OFF). In the first state, the inductor L1 and L2 are in charge energy phase, the second state inductor L1 and L2 releases the previously discharged energy (energy discharge) energy. The energy that has been released from the L1 inductor is used to charge the capacitor C1, while the energy from L2 is transferred to the load.

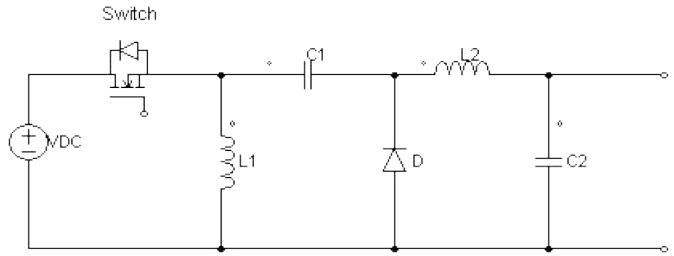


Fig. 8. Zeta Converter

$$\frac{V_o}{V_{in}} = \frac{D}{D-1} \quad (7)$$

$$L_1 = L_2 = \frac{Vin \times D}{\Delta IL \times Fs} \quad (8)$$

$$C_1 = \frac{Iout \times D}{Vout \times Fs} \quad (9)$$

$$C_2 = \frac{Iout \times D}{Vout \times 0.2 \times Fs} \quad (10)$$

Switching used is 40 KHz. By using high frequency the output signal from zeta converter will be smoother compared to buck-boost converter. Table I shows the specifications of the zeta converter used in the simulation according to calculations using equations 7 to 10.

TABLE I. PARAMETER SPECIFICATION AND DESIGN

Parameter	*Symbol*	*Value*	*Unit*
Input Voltage	Vin	21.2	V
Switching Frequency	f	40	KHz
Output Voltage	Vo	15	V
Rated Output Power	Po	100	W
Current Ripple	ΔIL	0.1	%
Voltage Ripple	ΔV_c	2	%
Inductor 1	L1	170	uH
Inductor 2	L2	170	uH
Capacitor 1	C1	1640	uF
Capacitor 2	C2	2300	uF

III. RESEARCH AND DISCUSSIONS

A. System Design

The proposed system is the search for MPP points using several different algorithms, namely P & O, First-order Differential and Newton Raphson. These three methods have the same level capabilities so that the most accurate and fast algorithm is obtained in the search for MPP points on solar panels. This MPPT uses a zeta converter. Solar panel specifications are shown in Table II.

TABLE II. SOLAR PANEL SPECIFICATIONS

Item	Value
Nominal Maximum Output (Pin)	100 W
Nominal Open Circuit Voltage (Voc)	21.2 V
Nominal Short Circuit Current (Isc)	6.46 A
Nominal Maximum Output Voltage (Vmpp)	17.1 V
Nominal Maximum Output Current (Impp)	5.84 A

Table II is a specification of the solar panels used in this paper. It is expected that the output power is always maximum 100 W with Vmpp 17.1 Volt and Impp 5.84A.

B. Simulation and Results

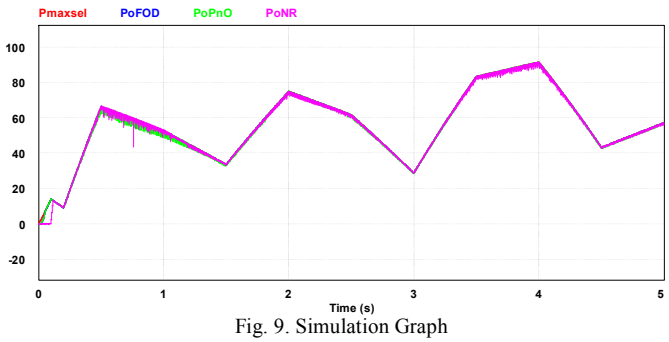


Fig. 9. Simulation Graph

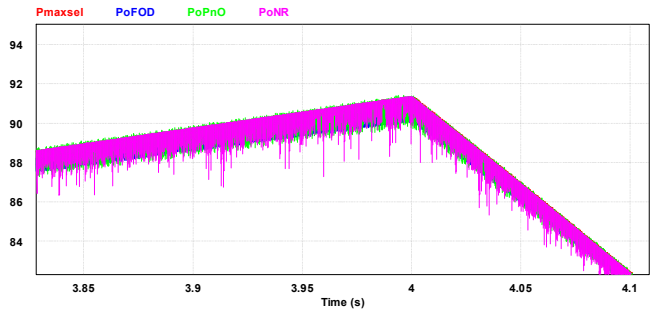


Figure 10. Graph of Simulation On Peak Shoot

The simulation data was taken by making irradiation changes 12 times and a constant temperature of 25ºC. Irradiation is given with varying values to determine whether the method applied is able to follow the changes in irradiation given. Figure 9 shows that the simulated First-order Differential, P & O and Newton Raphson method is able to follow the change of power according to the solar panel output. Figure 10 shows in detail the value of peak power from the use of these three methods. It can be seen that all three methods have different output power ripples. The P & O method has the largest output power ripple compared to the First-order Differential method and Newton Raphson. This can be affected by the value of the variable used to increase or decrease the value of the zeta converter duty cycle. With a large variable value, the ripple is large enough to be less accurate but fast in achieving maximum power. Whereas with a small variable value will be more accurate so that the power ripple is also small but requires a long time to reach the maximum power point of PV. The use of the First-order Differential and P & O methods has an output power value that is almost the same but lower compared to the Newton Raphson method. The similarity of power between the First-order Differential and P & O methods cannot be separated from the influence of the program between the two because in accordance with the flowchart program the First-order Differential and P & O methods are also almost the same. While the use of Newton Raphson method has the largest output power so that it approaches the maximum output power of solar panels. The output power values from the application of the P & O method, First-order Differential and Newton Raphson are presented in Table III.

TABLE III. COMPARISON OF POWER OUTPUT ACCURACY RATE OF PV

Irradiation	**The Power of PV**			
	The Maximal Power of PV	***First-order Differential***	***P&O***	***Newton Raphson***
0	0	0	0	0
100	95.906763	94.87394	95.76175	95.72824
300	29.38648	29.38684	29.41071	29.34592
350	34.0697	34.06901	33.92645	34.06944
450	43.35609	43.34881	43.2278	43.35646
550	52.308009	52.28332	52.20914	52.29478
600	56.96502	56.99589	56.99421	56.86321
650	61.65181	61.61213	61.51994	61.5755
700	65.60733	65.55677	65.39567	65.59921
800	74.35686	74.08333	74.25765	74.36466
900	83.26597	82.64179	83.33411	83.3357
1000	91.16	90.09459	90.52474	90.74816

TABLE IV. COMPARISON OF ERROR VALUE

Irradiation	**% Error Power of PV**		
	First-order Differential	***P&O***	***Newton Raphson***
0	0%	0%	0%
100	1.07690%	0.15120%	0.18614%
300	0.00123%	0.08245%	0.13802%
350	0.00203%	0.42046%	0.00076%
450	0.01679%	0.29590%	0.00085%
550	0.04720%	0.18901%	0.02529%
600	0.05419%	0.05124%	0.17872%
650	0.06436%	0.21389%	0.12378%
700	0.07706%	0.32262%	0.01238%
800	0.36786%	0.13342%	0.01049%
900	0.74962%	0.08183%	0.08374%
1000	1.16873%	0.69686%	0.45178%
Average %Error Power of PV	4%	0.26389%	0.12120%

Table III shows the output of the solar panel in each given irradiation. PV power is the maximum power of solar panels compared to PV power using the First-order Differential, P & O and Newton Raphson methods. Almost the same value (approaching) between PV power and power usage of these three methods shows that the use of the method is able to follow the maximum output power of PV. Table IV shows the error value of the power output obtained from the difference between the PV power output power and the use of the three methods. As long as the percent error data from the First-order Differential method is obtained 4%, P & O is 0.26389% and Newton Raphson is 0.12120%. So it can be seen that the use of Newton Raphson method has a smaller error compared to the First-order Differential and P & O methods.

IV. CONCLUSIONS

From the Comparative study the First-order Differential, Perturb and Observe and Newton Raphson Algorithm methods on MPPT-Zeta obtained PV power output data with the three methods applied which is close to the maximum value of PV output power. So that shows that all three methods can be used in finding Maximum Power Point Tracking. Among the First-order Differential, P & O and Newton Raphason methods each has advantages and disadvantages as described above. However, it can be concluded that the Newton Raphson method has the best accuracy and speed in achieving maximum power PV. This is intended by using the Newton Raphson method to have the power output that is closest to the maximum PV power with the smallest power error value compared to the First-order Differential and P & O methods. As with simulation data where the First-order Differential method has a power error of 4%, P & O is 0.26389 & Newton Raphson has only a power error of 0.12120%.

REFERENCES

[1] W.K. Francis, SS. Beevi, J. Mathew," MATLAB/Simulink PV Module Modelof P&O And DC Link CDC MPPT Algorithmswith Labview Real Time Monitoring And Control Over P&O Technique," International Journal of Advanced Research in Electrical, Electronics and Instrumentation
Engineering 2014.

[2] Logeswaran, T., & Kumar, A. "A Review of Maximum Power Point Tracking Algorithms for PV Systems under Uniform and Non-Uniform irradiances," Energy Procedia 54, 2014

[3] Soedibyo, Murdianto F.D, Suyanto, A. Mochamad, P. Ontoseno,"Modeling and simulation of mppt SEPIC combined bidirectional control invers KY converter using ANFIS in microgrid system," Indonesian Journal of Electrical Engineering and Computer Science, vol. 1, pp. 264-272, February 2016.

[4] Suyanto, Murdianto F.D, P. Ontoseno, S. Adi, S. Rony,"Wind-PV hybrid system modeling using bidirectional converter with MPPT-dual adaptive neuro fuzzy inference system (ANFIS) in microgrid isolated system," Journal of Engineering and Applied Sciences, vol. 11, pp. 2353-2359, December 2016.

[5] Trishan E., C. P., "Comparison of Photovoltaic Array Maximum Power Point Trackin Techniques," 35th Annual IEEE Power Electron 2007.

[6] Jee-Hoon Jung, and S. Ahmed, "Model Construction Single Crystalline Photovoltaic Panels For Real-Time Simulation," IEEE Energy Conversion Congress & Expo, September 12-16, 2010, Atlanta,USA

[7] M.C.D. Piazza, A. Ragusa and G. Vitale, "Identification of Photovoltaic Array Model Parameters by Robust Linear Regression Methods" ICREPQ, April 15-17, 2009, Valencia (Spain)

[8] N. Khaldi, H. Mahmoudi, M.Zazi and Y.Barradi, "The MPPT Control of PV System by Using Neural Netework Based on Newton Raphson Method" IEEE, 2014

[9] N. Sivanithra, K. Balamurugan, St. Michealcollage of Engineering and Technology Kallayarkovil, "Implementation of Zeta Converetr based Water Pump Systwm" ICET, 2017

[10] N. Karamia, "Analysis and Implementation of an adaptive PV based battery floating Charger", Solar Energy, Vol 86, pp. 2383–2396, 2012.

[11] M. Z. Efendi, F.D. Murdianto and R. E. Setiawan, "Modeling and Simulastion of MPPT Sepic Converter using Modified PSO to Overcome Partial Shading Impact on DC Microgrid System," International Electronics Symposium on Engineering Technology and Applications (IES-ETA), 2017

[12] F. D. Murdianto, A. R. Nansur, and A. S.. L. Hermawan, "Modeling and Simulation of MPPT Couple Inductor Sepic Converter using Flower Pollination Algoritm (FPA) Method in DC Microgrid," International Electronics Symposium on Engineering Technology and Applications (IES-ETA), 2017

[13] J. Falin "Designing DC/DC converters based on ZETA topology" Texaas Instruments Incorporated, Power Management, 2010

[14] S. H. Hoseini, A. Farakhor, and S. K. Haghighian, "Novel Algorithm of MPPT for PV Array Based on Variable Step Newton-Raphson Method Through Model Predictive Control" 13th International Conference on Control, Automation and Systems, Oct. 20-23, 2013, Gwangju, Korea.

[15] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.

[16] S. Amara, A. Bouallegue and A.Khedher, "Theoretical and Practical Study of a Photovoltaic MPPT Algorithm Applied to Voltage Battery Regulation," International Journal of Renewable Energy Research Vol.4, No 1, 2014

[17] P. Midya, P. T. Krein, R. J. Turnbull, R. Reppa, and J. Kimball, "Dynamic Maximum Power Point Tracker for Photovoltaic Aplications," 27th Annual IEEE Power Electronics Specialists Conference.

[18] S.H. Jubair,, K. Suhas, and A.Lokesh, "Development of Improved Curve Tracker for Photovoltaic System under partial Shaded Condions," International Conference on Circuit, Power and Computing Technologies [ICCPCT], 2014

[19] F. D. Murdianto, O. Penangsang, and A. Priyadi, "Modeling and simulation of mppt-bidirectional using adaptive neuro fuzzy inference system (ANFIS) in distributed energy generation system," IEEE International Seminar on Intelligent Technology and Its Applications (ISITIA), 2015.

[20] S. Chun and A. Kwasinski, "Analysis of Classical Root-Finding Methods Aplplied to Digital Maximum Power Point Tracking for Sustainable Photovoltaic Energy Generation," Ieee Transactions On Power Electronics, Vol. 26, No. 12, December 2011.

[21] S.P. Singh, A.K. Gautam, S.P. Tripathi, and B. Kumar, "Performance Camparison of MPPT Techiques using Cuk converter for Photovoltaic Energy Conversion System," IEEE-CICT, 2017

[22] T. Lodh and T. Majumder, "Highly Efficient Integrated Zeta-Flyback DC-DC Converter fot Hifgt Gain Aplication with a Compact Stucture," IEEE, 2016

[23] L. Qin and X.Lu, "Matlab/Simulink-Based Research on Maximum Power Point Tracking of Photovoltaic Generation," International Conference on Applied Physics and Industrial Engineering 2012

[24] F.D. Murdianto, M. Z. Efendi, R. E. Setiawan and A. S.. L. Hermawan, "Comparison method of MPSO, FPA, and GWO algorithm in MPPT SEPIC converter under dynamic partial shading condition", International Conference on Advanced Mechatronics, Intelligent Manufacture, and Industrial Automation (ICAMIMIA), 2017

[25] F.D. Murdianto, A. S.. L. Hermawan, A. R. Nansur and R. E. Setiawan, "Comparison method of flower pollination algorithm, modified particle swarm optimization and perturb & observe in MPPT coupled inductor sepic converter on DC microgrid isolated system", International Conference on Advanced Mechatronics, Intelligent Manufacture, and Industrial Automation (ICAMIMIA), 2017

[26] Rakhmawati R, Wahjono E, Suhariningsih, Murdianto F.D, Andriawan D,"Design and Implementation Boost Converter with Constant Voltage in Dynamic Load," International Seminar on Application for Technology of Information and Communication (iSemantic), 2017

Design and Realization of Low-Cost Wireless Remote Infusion Monitoring System

Agung W. Setiawan, Nurul Yenas, Deddy Welsan
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
awsetiawan@stei.itb.ac.id, nurulyenas@gmail.com, deddy.welsan@gmail.com

***Abstract*— One of the most frequently used in healthcare services is infusion systems. There are several types of infusion systems. The simplest version utilizes gravity and clamping device, the modern one uses the infusion pump to control the rate. The focus of this research is on the development of wireless remote monitoring infusion systems that can accommodate any infusion systems. In the last decade, the hospital has adopted a patient-centered care design. The consequences are the distance between the nurses' station and the patient room even further, and nurses spend more time to reach the patient room. So, it is important to develop remote infusion monitoring system to improve patient safety. This research has designed and developed a low-cost remote infusion monitoring system. The SWOT analysis is used to determine the features of the system. There are some important features that have been implemented, such as reporting the important information to the nurses' station; realization of the monitoring system in the patient room and nurses' station using wireless technology. This paper has investigated the FDA's recommendations on radio frequency wireless technology in medical devices and tries to implement it in Indonesia. The first one is the selection of wireless technology. From the study, the Wi-Fi module is chosen. The second one is the security of wireless signals and data. This recommendation is implemented using WPA. The last recommendation is wireless QoS. In this research, the MAC priority settings are used to implement the recommendation. WPA and MAC priority settings are set in the Wi-Fi access point.**

Keywords—wireless; remote; monitoring; medical; device

I. INTRODUCTION

One of the main concern of healthcare service today is patient safety. According to the World Health Organization (WHO), the definition of patient safety is the reduction of risk of unnecessary harm associated with healthcare to an acceptable minimum [1]. The trend of medical equipment technology development is on improving the safety of medical device to reduce and prevent the probability of medical errors [2]. The actions that can be taken to reduce the risk is to monitor the change in patient status using device alarm and make sure that this alert is heard and checked immediately by healthcare professionals.

Infusion systems are one of the most frequently used in healthcare services to deliver fluids or drugs to the patient. The simplest infusion systems use gravity and clamping device to control the flow of the liquid. The modern one is using an infusion pump. This paper focuses on the design and development of remote monitoring infusion systems that can be used by any type of infusion system device. So, this is a complementing device of the existing infusion systems device. Although there is an advanced infusion pump, it's still necessary to use this complementary device. In infusion systems, it's important to monitor the flow rate, the volume of the fluid and time of infusion. For more than a decade, hospital design has adopted the patient-centered care concept [3]. It makes the distance between patient room and nurses station even further, and nurses spend more time to reach the patient room. Due to this reason, it's important to implement remote monitoring, especially in infusion systems monitoring, so that the nurses can check the patient status immediately. The aim of this research is to improve patient safety through the design and development of remote infusion monitoring systems that can be used by any infusion systems devices.

II. DESIGN

To identify the research position and opportunities for developing infusion monitoring system; the Strengths, Weaknesses, Opportunities, and Threats (SWOT) analysis is used in this paper. The summarize of the analysis can be seen in Table 1.

TABLE I. SWOT ANALYSIS

Internal	**Strengths** • Real-time measurement: fluid rate and volume, • Remote wireless monitoring, • Can be used for any type of infusion system, • Alert system in patient room and nurses station.	**Weaknesses** • No pump and controller to adjust the infusion rate.
External	**Opportunities** • In 2016, there are 2,601 hospitals and 3,411 inpatient PUSKESMAS in Indonesia, • A lot of low-cost electronic components and modules that have quality and reliability, • Implementation of advanced technology to improve patient safety.	**Threats** • Several infusion systems have complex and more features, • More cost to invest in Wi-Fi installation network, if there is no existing computer network.

Several features in this infusion system that is developed to improve patient safety, there are:

- Fluid volume detection,
- Alert system, alarms in both patient room and nurses station if system detects running out of fluid (less than 15 ml),
- Amount of infusion fluid and the rate are displayed in the nurses station,
- Wireless communication systems for transmitting data from the device to nurses station to minimize hard-wired installation and connection,
- There are two power sources, AC power source and battery for power back-up.

According to the United States (U.S.) Food and Drug Administration (FDA), the developers of medical telemetry devices are encouraged to use three frequency bands of: 608-614 MHz, 1395-1400 MHz, and 1427-1432 MHz. These bands are allocated by The Federal Communications Commission (FCC) exclusively for wireless medical telemetry. In Indonesia, according to Decree of the Minister of Communication and Information Technology Number 25 Year 2014 on Table of Allocation of Radio Frequency Spectrum, the three frequency bands that are encouraged by FDA, 608-614 MHz; 1395-1400 MHz; and 1427-1432 MHz are already assigned for another use (not wireless medical telemetry use). Besides these bands, FDA also states that the medical telemetry devices can transmit in Industrial, Scientific, and Medical (ISM) bands: 900 MHz, 2.4 GHz, and 5.8 GHz. These are unlicensed basis frequency bands and defined by the International Telecommunication Union (ITU) Radio Regulations.

Another wireless technology module that can be used is Bluetooth and Zigbee. The IEEE 802.15, the working group of the Institute of Electrical and Electronics Engineers (IEEE) which specifies wireless personal area network (WPAN) standards, has announced will adopt Bluetooth as the IEEE 802.15 standard. This standard and Wi-Fi (IEEE 802.11) are complementary. Bluetooth has several limitations, class 2 Bluetooth device has relatively low transfer data rate, point-to-point data transmission, the range of transmission distance 0-10 meters. The class 2 Bluetooth device is unsuitable for transmitting large data over distance more than 3 meters in the presence of a wall barrier[4].

From the study on wireless technology module in the medical application, it can be concluded that the Wi-Fi module is more reliable for remote infusion monitoring systems, due to there is a wall barrier between patient room and nurses station. While Bluetooth is a point-to-point transmission, using Wi-Fi, the signal strength can be increased with Wi-Fi router.

III. Analysis

The common feature in infusion monitoring systems is fluid rate detection. This research has one important additional feature, the systems are equipped with a load cell sensor to detect the amount of the fluid in the infusion bag. This systems is part of the alert systems and introduced to minimize the human error and improve patient safety.

The block diagram of the proposed remote infusion monitoring systems is shown in Figure 1. The main part of the systems consists of two sensors: optical-based sensor and load cell, microcontroller, LCD displays, alarms, wireless communication module.

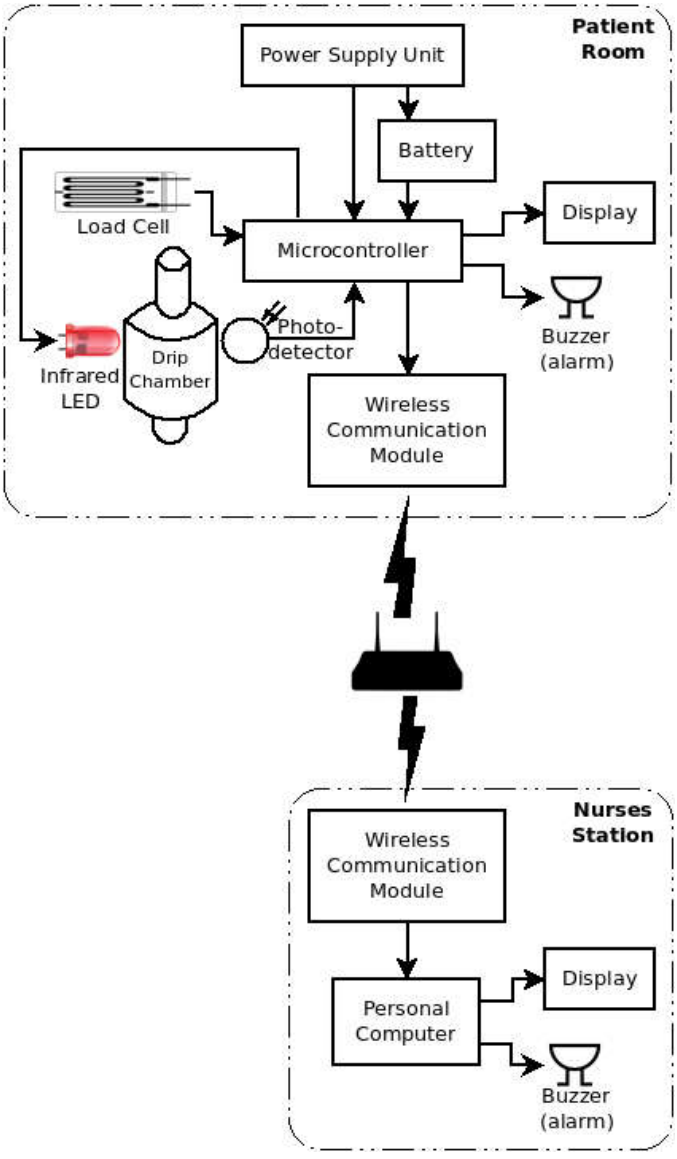


Fig. 1. The block diagram of the proposed remote infusion monitoring systems.

The infusion fluid rate is detected using an optical-based sensor that's consists of an infrared Light Emitting Diode (LED) and a photodiode. Infrared LED as the light source was choosen to minimize and eliminate disturbance from the environment, especially from the visible light lamp that's installed in the patient room. The detection is based on light

intensity, if there is a fluid drip, the light intensity will decrease. The intensity is lower than if there is no fluid drip. These intensities are detected using a photodiode. The function of the photodiode is to convert the light intensity into voltage. The outputs from these sensors are analog quantities. The Analog-to-Digital Converter (ADC) is used to convert these quantities to digital, so it can be processed by the microcontroller. The time interval between two conditions also can be measured. This rate detector is installed on the drip chamber.

To monitor the amount of fluid in the infusion bag, a load sensor is used. The assumption that is used in this research is the density of infusion fluid is the same as still water, 1 Kg/L = 1 gr/cc. Using this assumption, the fluid volume in the infusion bag can be estimated.

Another important module in this systems is a microcontroller board. The function of this module is to collect the data from the sensors, process the data, display and send the results. In this paper, Arduino Uno is used, due to it has several input/output pins to accommodate the proposed design and suitable for prototyping. For the further development, next version, the selection of microcontroller needs to be optimized based on price and technical specification consideration; the number of input-output pins, RAM, and Analog-to-Digital Converter (ADC).

Using the time intervals and the volume of the infusion fluid that's still remained in the infusion bag, the infusion rate can be calculated. The mathematical formula of this calculation is processed in the microcontroller. Liquid-crystal display (LCD) matrix is used to show the infusion rate and fluid volume that is remained in the infusion bag.

Besides the results is displayed in the patient room, the main feature of the systems is remote monitoring. The results are sent to the nurses' station through the Wi-Fi connection. The data that's is sent not only these results, but also room number, patient ID, and device ID. So the healthcare professionals can monitor the condition of the patients.

The system is equipped with the alarms. This is an alert system to improve patient safety, reduce and prevent the probability of medical errors. The wrong rate is the most frequent serious errors, another error is related to the volume[5]. There are two buzzers in this systems that are attached to the device and installed on the nurses' station. If the systems detect unnormal conditions, both alarms will be ringing. So the healthcare professionals can response this immediately.

In this research, the ESP8266 is used to implement the wireless connection between patient room and nurses station. There are some considerations to choose this Wi-Fi module. This is a low-cost Wi-Fi module and compatible with Arduino. In [6] FDA recommended that wireless medical devices deploy wireless protection. This research proposes two types of the wireless security. The first one is Wi-Fi Protected Access (WPA) for wireless data encryption. The second one is Media Access Control (MAC) address filtering that is set on the Wi-Fi access point. Only the device with a registered MAC address is granted to access the Wi-Fi network. Another alternative to implementing the wireless security is using Remote Authentication Dial-In User Service (RADIUS). This alternative did not use in this research due to this realization needs a RADIUS server, and it will increase the implementation cost.

The data from the device is sent to the nurses's room using wireless communication systems. FDA recommends that the healthcare facility should consider the Quality of Service (QoS). To accommodate the recommendations, this research use MAC priority settings in the access point. The registered MAC addresses will have higher priority access. So, the MAC address of the Wi-Fi module of the device should be registered to Wi-Fi access point.

To reduce the cost, the system is designed to use the Personal Computer (PC) in the nurses station. The assumption is nurses station is equipped with a PC that has a Wireless Local Area Network (WLAN) adapter. The PC monitor is used to display the data. The internal PC speaker is utilized as alarms.

In the volume of infusion fluid measurement, the first step is to get the weight of the fluid using a load cell sensor. It is important to calibrate the load before implemented to the systems. The sampling rate of this process is one sample per second. Then, the last 10 will be averaged to get the weight of the infusion fluid. To prevent measurement error due to shakes on the load sensor or IV container, an algorithm is applied to filter the measured value so that values that have too far deviation with the counted average value will be discarded. The volume value will be displayed to the LCD and sent to the nurses' station.

IV. Concluding Remarks

This research has designed and developed a low-cost remote infusion monitoring system. The SWOT analysis is used to determine the features of the system. There are some important features that have been implemented, such as reporting the infusion rate and the fluid volume that is still remained in infusion bag to the nurses's station; realization of the alert system in patient room and nurses' station; realization of the radio frequency wireless technology in this system. This paper has investigated that the radio frequency wireless technology can be implemented for remote infusion monitoring systems. According to the FDA's recommendation, this research has accommodate three considerations on the integration of radio frequency wireless technology into medical systems. The first one is the selection of wireless technology. From the study on wireless technology module in the medical application, it can be concluded that the Wi-Fi module can be used for wireless medical telemetry systems in Indonesia. For that reason, the Wi-Fi module is implemented in this research. Another FDA's recommendation that is addressed in this research is the security of wireless signal and data. The last recommendation is the wireless quality of service. These last two recommendations are set in the Wi-Fi access point configuration. Currently, this research has developed only one device. Future work will focus on the management information systems, to handle more than one device. So, it is important to design the information systems properly. Beside desktop-based

application, to accommodate mobility of the healthcare professionals, this research also plan to develop an Android-based application.

REFERENCES

[1] World Health Organization, Reporting and Learning Systems for Medication Errors: the Role of Pharmacovigilance Centres. Geneva, Switzerland: WHO Press, 2014.

[2] T.L. Rodziewicz, and J.E. Hipskind, Medical Error Prevention. Treasure Island (FL): StatPearls Publishing, 2018.

[3] H-B. Seo, Y-S. Choi, and C. Zimring, "Impact of Hospital Unit Design for Patient-Centered Care on Nurses' Behavior," Environment and Behavior, vol. 43(4), pp. 443-468, 2010.

[4] S. Pudaruth, H.K. Ramdolin, and A. Bissoonee, "An Assessment Of The Performance Of Bluetooth As A Broadcasting Channel," Proceedings of the World Congress on Engineering 2010, Vol I, pp.785-790, July 2010.

[5] J.I. Westbrook, M.I. Rob, A. Woods, and D. Parry, "Errors in the administration of intravenous medications in hospital and the role of correct procedures and nurse experience," BMJ Quality & Safety, Vol. 20(12), pp.1027-1034, 2011.

[6] U.S. Department of Health and Human Services Food and Drug Administration Center for Devices and Radiological Health, Radio Frequency Wireless Technology in Medical Devices: Guidance for Industry and Food and Drug Administration Staff. Rockville, MD, 2013.

Speed Identification As Vehicle Safety Feature Based Eulerian Color Magnification

Septian Enggar Sukmana
Informatics
Universitas Dian Nuswantoro
Semarang, Indonesia
septian.enggar@dsn.dinus.ac.id

Farah Zakiyah Rahmanti
Informatics
Universitas Dian Nuswantoro
Semarang, Indonesia
farah.zakiyah@dsn.dinus.ac.id

Mauridhi Hery Purnomo
Computer Engineering Dept.
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
hery@ee.its.ac.id

***Abstract*—Road accident which is caused vehicle collision is still big problem in Indonesia. Beside that, research of smart vehicle is still limit because not many supports to do it. Based on these two problems, the development of smart vehicle which exploits camera as collision detection to avoid accident is quite challenging. Speed identification may be a one of collision detection feature. Objective of this paper is to implement color magnification to vehicle object in video to improve speed identification result which is useful for safety feature in smart vehicle development. However, acquiring of speed data accuracy is still a challenging in researchers. Implementation of color magnification to vehicle objects that are captured by smart vehicle camera is an interesting concept that needs to be performed. Although there comes some different speed value that resulted from color magnification and non-color magnification implementation, it still be a compromise technique to achieve better result with some improvements in the next activity.**

Keywords—speed identification, smart vehicle, color magnification based eulerian, road accident, vehicle collision

I. INTRODUCTION (*HEADING 1*)

Road accident is still one of main problem in Indonesia, the number of this case has not decreased with average of 20,000 cases in the latest trimester [5]. Human factor is also the main factor, Kementerian Perhubungan Republik Indonesia has placed this factor to be the top factor of road accidental cause [4]. It is a serious problem because any abnormal movement by human can threat the other people in the road [9] [10]. When this problem still has to be solved manually, it comes to be a problem too for smart vehicle develoment because smart vehicle needs to generate reflection that is caused by another vehicle movement. Smart vehicle development is a new challenging in Indonesia and because of limitless support and funding to performs its research, more efficient technique, hardware, and software can be great challenge to smart vehicle researcher in Indonesia.

The awareness in development of smart vehicle needs special hardware and certain compromised algorithm in software. This awareness can be obtained by camera technology with special computational feature, Hui et.al [6] develops an algorithm to spot some vehicles that are captured by camera and detects collision when it happens. Example of implementation this feature to vehicle is Bernini et.al [3] who have exploited some computer vision techniques to optimize collision detection feature.

However, to detect and predict a collision it needs a speed identification to function its feature beside distance measurment between self-vehicle and its neighbour. Speed identification may help collision detection more accurate, especially on fast moving object. Combination between distance approximation and vehicle proximity would be a better step to reach the speed identification of vehicle. But, this processing still performs on spatial domain which is frequency-based domain may be a great challenge to achive a better result. One technique to do it is using color magnification to neighbour vehicle of smart vehicle that is captured by its camera. Actually, color magnification follows subtle motion of object which can be revealed by explicit or implicit movement estimation [1], but explicit movement will result more computation resource because needing of computation is required more than implicit one [1]. The spatio-temporal processing in video data which is exploited in implicit movement estimation does not burden to accuration of movement estimation. Lagrangian [7] [8] and Eulerian [1] [2] series is quite known as a technique to achieve subtle motion via implicit estimation. Form of color magnification is color variation of each object that be captured in video, so the movement of objects can be known in subtle motion visually. This is an advantage to retrive information of vehicle movement and it can be useful parameter to get better accuration of vehicle speed.

This paper presents implementation of color magnification in vehicle speed identification on road as development of smart vehicle feature via camera that may be implemented in Indonesia to avoid road accident. Color magnification technique will be implemented in each vehicle object that is captured by camera that be placed in smart vehicle. This is only limit in vehicle to vehicle, not vehicle to non-vehicle.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

II. METHODS

In this activity, there are two main activities such: color magnification and speed identification. Color magnification is performed based on eulerian video magnification (EVM) [1] and speed identification is performed by exploiting Manhattan distance to check proximity in vehicle. Video processing starts from EVM to make color magnification version of video, then speed identification runs the color magnification video version to get speed information.

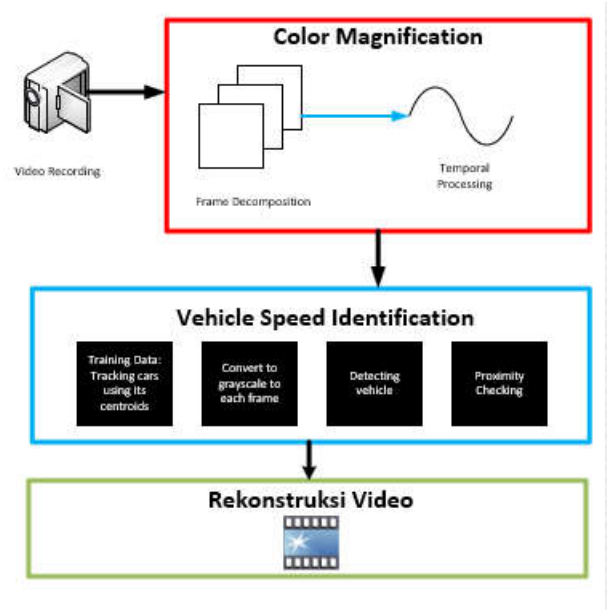


Fig. 1. Steps of acquiring data and implementation of color magnification in speed identification

Fig. 1 shows how this scheme works. Video recording activity is held by placing an action camera. The video data that resulted from this activity will performs EVM. This video will take spatial decomposition on each frame. Temporal processing may result signal synthesization with the same α magnification factor.

$$I(x,t) = f(x + (1+\propto)\delta(t)) \quad (1)$$

where *I(x,t)* is image intensity at position *x* and time *t*.

When first order Taylor expansion is imlemented, the (1) would change to

$$I'(x,t) = f(x) + \delta(t)\frac{\delta f(x)}{\partial x} \quad (2)$$

Broadband temporal bandpass filter has changed the image intensity

$$B(x,t) = \delta(t)\frac{\delta f(x)}{\partial x} \quad (3)$$

As we know that has important role, so this amplification also changes image intensity

$$I'(x,t) = I(x,t) + \propto B(x,t) \quad (4)$$

Because of this, the value of $I'(x,t)$ approximates to $f(x + (1+\propto)\delta(t))$. It is not a surprise because it happens from combination between (2), (3), and (4).

$$I'(x,t) \approx f(x) + (1+\propto)\delta(t)\frac{\delta f(x)}{\partial x} \quad (5)$$

It is not perfect yet, Taylor approximation may result some failures in image function, so $I'(x,t)$ must be equal to magnified function.

$$f(x) + (1+\propto)\delta(t)\frac{\delta f(x)}{\partial x} \approx f(x + (1+\propto)\delta(t)) \quad (6)$$

$f(x) = \cos \omega x$ is for spatial frequency ω and $\beta = 1 + \alpha$ for (7) then use addition law for cosines (8)

$$\cos(\omega x) - \beta\omega\delta(t)\sin(\omega x) \approx \cos(\omega x + \beta\omega\delta(t)) \quad (7)$$

$$\cos(\omega x) - \beta\omega\delta(t)\sin(\omega x) = \cos(\omega x)\cos(\beta\omega\delta(t)) - \sin(\omega x)\sin(\beta\omega\delta(t)) \quad (8)$$

Each frame has been processed by color magnification, so each color object that captured in the scene frame will be magnified. It is the new frame with color appearance as magnification result. The change position of color in sequence frame will be exploited to acquire speed information. In this acticity, to make program know what object woud be identified, some training data which consist some feature of every vehicle that consists in Indonesia will be utilised.

The new frame is testing data with converting frame to grayscale is first step, then vehicle detection is performed. Each object that has been detected is enumerated in centroid list. Centroid list contains some feature such: last updated frame, color, object position, count of object that has been locked, count of object is unlock, lock object by default, and list of object speed in previous frames. To check the proximity, Manhattan distance is used. To tackle the identification of closer vehicle object is solved by certain pixel proximity value of rectangle of previous frame assumption.

Because of color magnification has changed frame display condition, so it needs to evaluate this result. In this activity, PSNR (9) is used to evaluate in image quality after color magnification process and SSIM (10) is used to evaluate condition of frame after color magnification that is compared with original video frame.

$$PSNR = 10\ log_{10}\left(\frac{MAX_I^2}{MSE}\right) \quad (9)$$

$$SSIM\ (x,y) = [l(x,y)^{\alpha}, c(x,y)^{\beta}, s(x,y)^{\gamma}] \quad (10)$$

where MAX is maximum possible pixel value of image, *l* is luminance component, *c* is contrast component, and *s* is structure component. Last step is video reconstruction, it will assemble all frames into one single video again.

III. RESULTS

Collecting data was conducted by 50 times of video recording activity. After several trials, the best scene is the morning one because it is very difficult to get better result when night time or day time video is performed. Although speed measurement is depended on certain color magnification result which appears for short time, it is very compromize to get speed

information of the object. By default, the color magnification result will show variation color in the road environment. Fig. 3 shows result of color magnification, the frame on result has color appearance in environment scene. Because sensitiveness of color magnification may happen in certain video, the α value can be arranged manually to set it properly. In this data, we use α=150 as best value to perform color magnification.

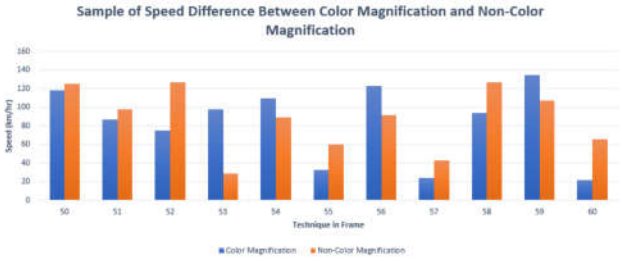
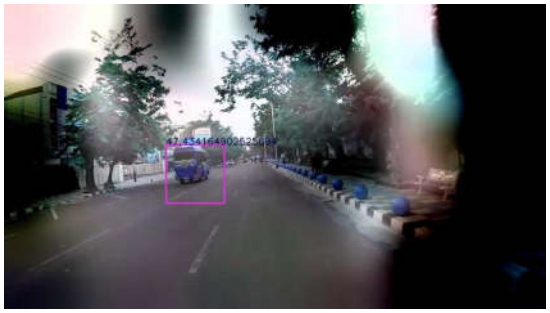


Fig. 2. Result comparison between color magnification and non-color magnification use

Fig. 2 shows speed value comparation between color magnification use and non-color magnification. It shows that there is still big gap value. By sample 50^{th} – 60^{th} frame as sample captured that includes two vehicles which is one is ours and another is not shows some frames have big different gap speed value especially on 52^{nd}, 53^{rd}, and 60^{th} frames although there is small gap speed value in 50^{th} dan 51^{st} frame.



(a)



(b)

Fig. 3. Speed identification result visually: a) without color magnification; b) with color magnification

Based on experimental activity, computational anlysis result shows no abberation because the height of pyramids does not have unconsistent different range value. Table 1 shows computational analysis result of five video sample in color magnification.

TABLE I. FIVE VIDEO SAMPLE'S COMPUTATIONAL RESULT OF COLOR MAGNIFICATION

Video -th	Frame Rate	Number of Frames	Height of Pyramids
1	55	854	15
2	53	888	15
3	55	933	15
4	54	787	14
5	55	650	15

Each video has different result in PSNR's frame. But, all of results have significant value of PSNR, there is no a little difference in each frame, it seems that color magnification gives intervention in image quality (Fig. 4). Increasing and decreasing qualities are also caused by color appearance.

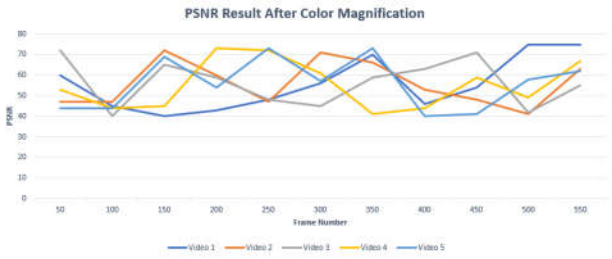


Fig. 4. PSNR result at each frame in each five videos sample

SSIM result has a few of high quality measure value. Fig. 5 shows SSIM result condition, it seems that some frames have little result value because color appearance affects SSIM mapping.

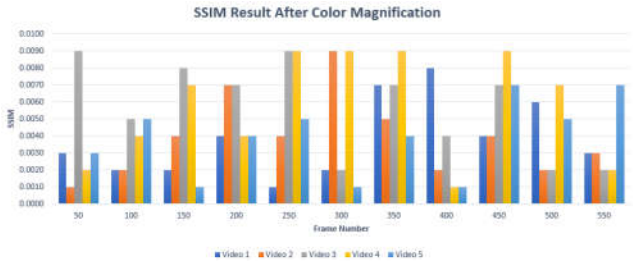


Fig. 5. SSIM result at each frame in each five videos sample

Color magnification implementation to vehicle speed identification results an unreal value of speed. Fig. 3 shows a frame sample result of color magnification implementation to vehicle speed identification, by comparing to non-color magnification video, it shows that result on color magnification

one gives more focused identification result (Fig. 3 (b)), it is different from Fig. 3 (a) that more results have been given although non-vehicle object is detected. But for more reliable result, based on those two frame results it is known that original frame is better that color magnification one because actually vehicle detection on original one detects more vehicles than color magnification one as impact of color magnification on frame that may cover vehicle object.

IV. CONCLUSION

The improvement of color magnification technique to this activity is still needed, it shows on big different value gap between speed identification with color magnification result and speed identification without color magnification result in some frames. Visually, color magnification still does not satisfy to make better accuration of speed value because some vehicles are identified has quite fast speed value (above 140 km/hr) and this phenomenon is not normal in Indonesia. The next activity will be use some robust detection on vehicle to reduce non-vehicle object be detected and color magnification with limited object area.

ACKNOWLEDGMENT

Authors would like to thanks to Kementerian Riset, Teknologi, and Pendidikan Tinggi Republik Indonesia for its support and research funding.

REFERENCES

[1] Wu H.Y, Rubinstein M, Shih E, Guttag J, Durand F, Freeman W. Eulerian Video Magnification for Revealing Subtle Changes in the World. *ACM Transactions on Graphics-Siggraph 2012*. 2012.

[2] Wadhwa N, Rubinstein M, Shih E, Guttag J, Durand F, Freeman W. Phase-Based Video Motion Processing. *ACM Transactions on Graphics-Siggraph 2013*. 2013.

[3] Bernini, Nicola. Bertozzi, Masimo. Castangia, Luca. Patander, Marco. Sabbatelli, Mario. 2014. Real-time obstacle detection using stereo vision for autonomous ground vehicles: A survey. 2014 IEEE 17th International Conference on Intelligent Transportation Systems (ITSC). halaman 873-878.

[4] Komite Nasional Keselamatan Transportasi Kementrian Perhubungan Republik Indonesia. 2016. *Data Investigasi Kecelakaan LLAJ Tahun 2010-2016*.Jakarta.

[5] Korps Lalulintas Kepolisian Republik Indonesia. 2017. *Kecelakaan di Indonesia selama triwulan terakhir*.URL: http://www.korlantas-irsms.info/graph/accidentData?lang=id (accessed: 19 Juni 2017, 23: 11 WIB)

[6] Hui, Zu. Yaouhua, Xie. Lu, Ma. Jiansheng, Fu. 2014. Vision-based real-time traffic accident detection. Proceeding of the 11th World Congress on Intelligent Control and Automation. pages 1035-1038. Shenyang, China.

[7] Wang J, Drucker S.M, Agrawala M, Cohen M.F. The Cartoon Animation Filter. *ACM Transactions on Graphics-Siggraph 2006*. 2006. 25: 1169-1173.

[8] Liu C, Torralba A, Freeman W.T, Durand F, Adelson E.H. Motion Magnification. *ACM Transaction on Graphics*. 2005. 24: 512-526.

[9] Masilkova, Monika. 2017. Health and social consequences of road traffic accidents. Social Science in Health, Article in Press.

[10] Porada, V. 2000. Silnicni dopravni nehoda v teorii a praxi. Linde. Praha.

Reverse Polarity Scanning Method on 64-Channel Capacitive Sensor to Improve the Performance of ECVT System

Arbai Yusuf[1,2], Agus Santoso Tamsir[1], Dodi Sudiana[1], and Harry Sudibyo S.[1]

[1]Department of Electrical Engineering, Universitas Indonesia, Depok, Indonesia
[2]CTECH Labs Edwar Technology Co, Tangerang, Indonesia
Email: arbai.yusuf41@ui.ac.id; dodi.sudiana@ui.ac.id; tamsir@ee.ui.ac.id; harisudi@eng.ui.ac.id

***Abstract*—** This research proposes a scanning method on tomography namely electrical capacitance volume tomography (ECVT). This study discusses the new scanning method of reverse polarity scanning that is applied to 64-channel capacitive sensors. In tomography process, the data acquisition system measures the capacitance from capacitive sensors using scanning process by activating the electrode pairs. Generally, scanning process in ECVT uses conventional scanning by activating an electrode pair and measure the capacitance signal for the pair. This study proposes a new scanning process, by reversing the polarity of electrode in scanning method. The method is applied by changing the polarity of excitation electrode into detection electrode, and vice versa. Then, both of methods are compared to evaluate the performance of the acquisition system. Simulation and experiment are performed using 64-channel capacitive sensors with cylinder rods phantom filled with water. To evaluate the performance of the ECVT system, static objects were used to measure parameters such as Mean Absolute Error (MAE), Coefficient-Correlation (R), and image qualitative observation. From the simulation and experiment, the Coefficient-*Correlation* (R) value as 0.6 and *Mean Absolute Error* (MAE) value as 0.2, so that the proposed method can improve image quality of 3.2%.

Keywords—ECVT; capacitive sensor; reverse polarity scanning; stray capacitance, ill posed.

I. INTRODUCTION

Visualization of objects in a vessel based on the measurement of capacitance is called capacitance tomography. The principle of capacitance tomography is by injecting a sinusoidal or square wave signal with a certain frequency to one of the electrode pairs as shown in Fig. 1. As a result, the signal will generates a fringe electric field inside sensor which will affect the object in the sensor [1]. Subsequent, the signal changes are detected by the other electrodes to be measured by the signal detection circuit. Then the signal is processed to be returned to an object using a specific reconstruction algorithm. There are two types of capacitance tomography i.e. ECT (electrical capacitance tomography) and ECVT (electrical capacitance volume tomography). The difference is, the design of sensors in ECT generally use a layer with a maximum number of electrodes 16 pieces. While the design of sensors in ECVT using multiple layers with the number of electrodes more than 24 pieces. The capacitance measurement in the capacitive sensor is carried out directly from the first level electrode to the last level electrode, then the data is reconstructed using a volumetric reconstruction algorithm according to the designed sensor. The result is a volumetric image that corresponds to the original form, and therefore namely ECVT tomography.

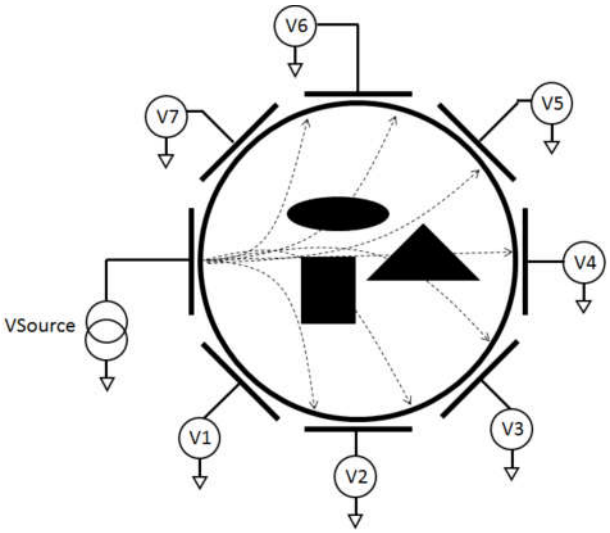


Fig 1. Principle of capacitance tomography

Many ECVT applications are used for multiphase flow [2], [3], gaseous and solid imaging [4]. In recent applications, ECVT has been applied to diagnose abnormalities in biological tissues, such as breast and brain [5]–[7]. The problem often encountered in capacitance tomography is image resolution is still low so that the image reconstruction result is more artifacts. As a consequence, the generated image seems to be directed toward the stronger sites of the sensitivity. Some researchers have been tried to improve the resolution and sensitivity of ECVT measurements, among others by arranging the sensitivity map distribution in capacitive sensors [8], adjusting the electrode configuration [9], and adjusting the electrode rotation [10].

Each electrode arrangement and size are made as accurate as possible by the distance between the electrodes, and the distance between the screen guards is made equal. Screen guard is a shielding made of aluminum or copper foil that connected to ground where its function is to protect the sensor from

electromagnetic field interference. However, in the application there are still differences in dimensions of each electrode size and errors in fabrication. In addition, each electrode is connected to one signal generator and one signal detection circuit, so that each electrode has its own characteristics. In the reconstruction algorithm, objects are regarded as non-isotropic which have different values if measured in different directions. Accordingly, the measurement of the capacitance of a pair of electrodes will differ in value if the polarity of electrode pair is reversed.

To overcome the above issue, this research proposed new scanning method that is reverse polarity scanning method. The measurement method is by reversing the polarity of each active electrode by changing the excitation electrode into detection and vice versa. With this method, the amount of measurement data becomes twice as much as the conventional method, thus reducing the main problem in the ill-posed reconstruction algorithm. Ill-posed is the least amount of measurement data when compared to the number of known voxel values causing the inverse S-matrix value unknown.

II. 64-Channels Capacitive Sensor

In this study designed cylinder-shaped sensor consisting of 64 electrodes with 8 levels, where each level has 8 electrodes as shown in Fig. 2. The sensor has a height of 24 cm, a diameter of 9 cm, a sensor material made from pvc (polyvinyl chloride), and 3x3 cm electrode size. Electrostatic can damage the electronic circuit used to measure capacitance in data acquisition systems. To overcome the problem, each electrode added 1 Mohm resistor which is connected with earth guard [11].

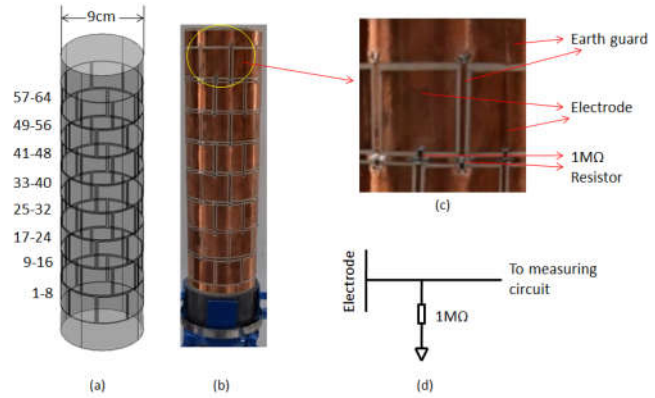


Fig 2. 64-channels cylindrical ECVT sensor; (a) sensor design (b) photograph of the sensor; (c) electrode connection in detail (d) circuit diagram of each electrode

In accordance with the principle of measuring capacitance, two parallel and opposite electrodes will have capacitance values. The value depend on the shape of geometry of the electrode, area of the electrode, distance between the electrodes, and the insulating material (dielectric) between the two electrodes. Analysis of the sensor design shows that between the electrode and the earth guard will give a capacitance value, where in tomography namely stray capacitance. Ideally, the stray capacitance of the sensor can be expressed in Eq. (1) as follows:

$$C_s = NS\frac{\varepsilon_0 \varepsilon_r A}{d} \quad (1)$$

Where, C_s is stray capacitance in the sensor, N is number of electrodes, S is number of sides in a single electrode, ε_0 is dielectric constant in a vacuum of 8.85 pFm^{-1}, ε_r is the relative permittivity of the dielectric material, A is area of electrode, and d is the distance among electrode and earth guard. If the parameters mentioned above are fed to Eq. (1), the stray capacitance of the sensor is 0.0177 pF for one electrode and 4.53 pF for 64 electrodes.

III. ECVT Scanning Method

Scanning is a way of obtaining data from the capacitive sensor, where the number of data depends on the number of electrodes and the scanning method. There are several methods in the ECVT scan, such as conventional scanning which is a common method. Another method is to arrange the electrode configuration, which combines two electrodes into one excitation and detection electrode [9], and rotary electrode [10]. In this study described conventional multistage scanning applied to the 64-channel ECVT systems.

A. Conventional Scanning Method

Generally the scanning method used in the ECVT system is conventional method, viz. activating electrode pair and measure the capacitance signal, hence the amount of data can be known by Eq. (2).

$$N = n(n-1)/2 \quad (2)$$

Where, N is number of capacitance data, n is number of electrode. For example, the sensor has 8 electrodes then the total capacitance data is 28 data. The conventional scanning process is shown in Fig. 3 through 5. In this example, the number of electrodes is determined by 8 pieces to simplify the explanation. The electrode is marked in different color, the red color is excitation electrode while the green color is detection electrode. There are seven scanning processes that start from the first excitation electrode and end with the seventh excitation electrode.

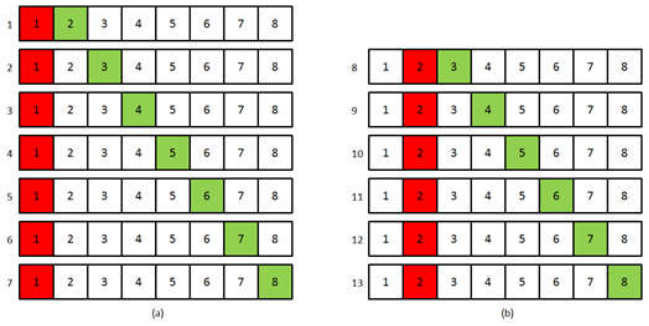


Fig 3. ECVT conventional scanning method; (a) excitation electrode number 1; (b) excitation electrode number 2

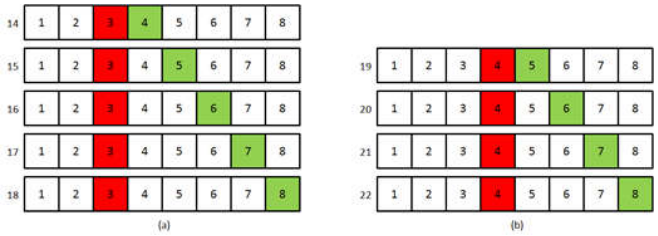


Fig 4. ECVT conventional scanning method; (a) excitation electrode number 3; (b) excitation electrode number 4

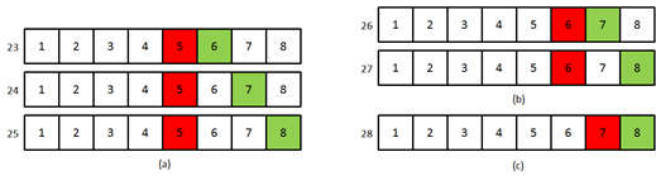


Fig 5. ECVT conventional scanning method; (a) excitation electrode number 5; (b) excitation electrode number 5; (c) excitation electrode number 7

The scanning process starts from the excitation electrode of number 1 and detection electrode number 2 to obtain the first data, the next scan for detection electrode number 3 to obtain the second data, and so on until the eighth detection electrode to obtain the seventh data. After the final detection electrode is scanned, the scan proceeds by activating the second excitation electrode and the third detection electrode to obtain the eighth data, and so on up to the eighth electrode and thirteen data is obtained. After the final detection electrode is scanned, the scan is continued with the next excitation electrode, so on until all electrodes are scanned with the result of 28 data. Table 1 describes the data acquisition of the scanning process in the ECVT system.

TABLE I. DATA ACQUISITION OF THE SCANNING PROCESS

Detection Electrode / Excitation Electrode	1	2	3	4	5	n
1	-	$C_{1,2}$	$C_{1,3}$	$C_{1,4}$	$C_{1,5}$	$C_{1,n}$
2	-	-	$C_{2,3}$	$C_{2,4}$	$C_{2,5}$	$C_{2,n}$
3	-	-	-	$C_{3,4}$	$C_{3,5}$	$C_{3,n}$
4	-	-	-	-	$C_{4,5}$	$C_{4,n}$
5	-	-	-	-	-	$C_{5,n}$
n	-	-	-	-	-	-

According to the Table 1, for example the number of electrodes is known to be 8 pieces, the capacitance readings are obtained from electrode pairs 1-2, 1-3, .., 1-8, 2-3, 2-4, .., 2-8, 3 -4, 3-5, 3-6, .., 3-8, and so on until the pair of 7-8 electrodes. If there are a number of n electrode, then the data acquisition of the capacitance corresponds to Eq. (1).

B. Reverse Polarity Scanning Method

The scanning method proposed in this study is reverse polarity scanning method. The scanning method is essentially similar to conventional scanning, except the measurement of an active electrode pair is taken twice. The method of measurement is by reversing the polarity of each active electrode i.e. changing the excitation electrode into detection and vice versa. Thus the acquisition of the capacitance data can be known by Eq. (3).

$$N = n(n-1) \qquad (3)$$

Where N is number of data measurement, n is number of electrode. For example, the sensor with the number of electrodes of 8 pieces, then the total data measurement is 56. Fig. 6 to Fig. 9 describes the scanning process using reverse polarity scanning method. In the picture, given an example of 8 pieces of electrodes to facilitate the explanation. As an illustration, the excitation electrode is colored red while the detection electrode is colored green. There are eight scanning processes starting from the first excitation electrode and ending with the eighth excitation electrode.

In the first process the excitation electrode number 1 and the number 2 are activated to obtain the first data, then the scanning goes up to the detection electrode number 3 and so on until the detection electrode number 8, hence one scanning process are completed and 7 data are obtained. The next process for the 2nd excitation electrode is activated, then the detection electrode is activated starting from electrode number 1, electrode number 3, electrode number 4, and so on until the number 8 electrode to obtain 7 data. The next process for the excitation electrode number 3 is activated and the detection electrode is activated starting from electrode number 1, electrode number 2, electrode number 4, and so on up to the number 8 electrode to obtain 7 data. The process is repeated until the eighth excitation electrode, so that the overall data acquisition of 56 capacitance data are obtained. The difference with conventional scanning is on conventional scanning, in one active excitation electrode the data is only taken once on the next electrode. While in the reverse polarity scanning method, in one active excitation electrode the data is taken twice for the previous electrode and the subsequent electrode.

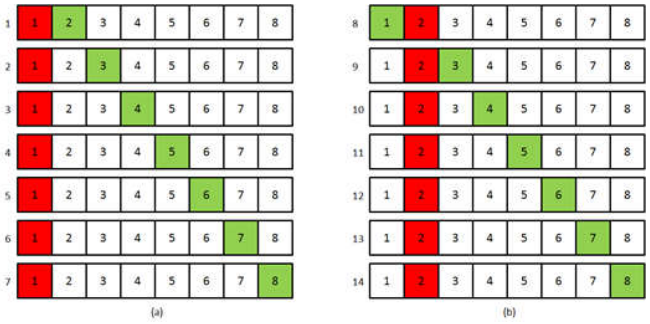


Fig 6. ECVT reverse polarity scanning method; (a) excitation electrode number 1; (b) excitation electrode number 2



Fig 7. ECVT reverse polarity scanning method; (a) excitation electrode number 3; (b) excitation electrode number 4

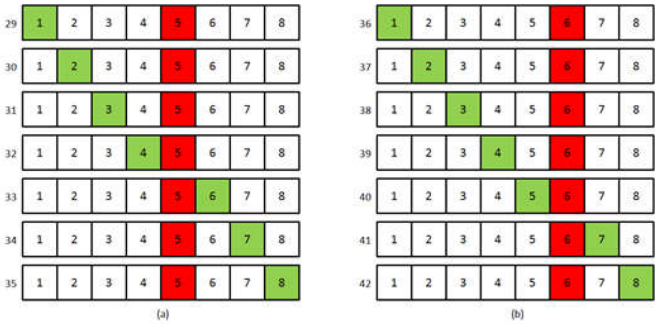


Fig 8. ECVT reverse polarity scanning method; (a) excitation electrode number 5; (b) excitation electrode number 6

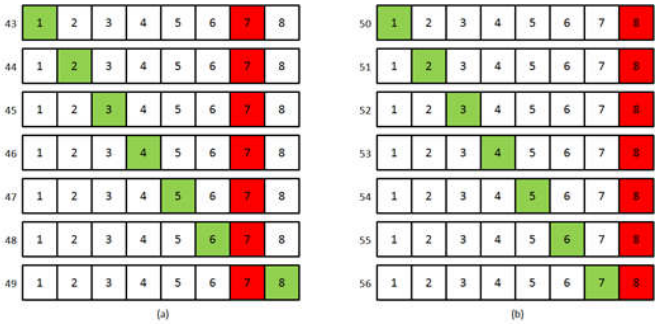


Fig 9. ECVT reverse polarity scanning method; (a) excitation electrode number 7; (b) excitation electrode number 8

The data acquisition of the reverse polarity scanning method is shown in Table 2. For example the number of electrodes of 8 pieces then the readings of the capacitance by the reverse polarity scanning method are obtained from electrode pair 1-2, 1-3, .., 1-8, 2-3, 2 -4, .., 2-24, 3-4, 3-5, 3-6, .., 3-8, and so on up to 7-8. Then the readings are repeated for different polarities i.e. 2-1, 3-1, ..., 8-1, 3-2, 4-2, ..., 8-2, 4-3, 5-3, 6-3, ..., 8-3, and so on up to 8-7. If there are electrodes of a number of n pieces then the acquisition of capacitance data is in accordance with Eq. 3 above.

TABLE II. DATA ACQUISITION OF THE SCANNING PROCESS

Detection Electrode / Excitation Electrode	1	2	3	4	5	n
1	-	$C_{1,2}$	$C_{1,3}$	$C_{1,4}$	$C_{1,5}$	$C_{1,n}$
2	$C_{2,1}$	-	$C_{2,3}$	$C_{2,4}$	$C_{2,5}$	$C_{2,n}$
3	$C_{3,1}$	$C_{3,2}$	-	$C_{3,4}$	$C_{3,5}$	$C_{3,n}$
4	$C_{4,1}$	$C_{4,2}$	$C_{4,3}$	-	$C_{4,5}$	$C_{4,n}$
5	$C_{5,1}$	$C_{5,2}$	$C_{5,3}$	$C_{5,4}$	-	$C_{5,n}$
n	$C_{n,1}$	$C_{n,2}$	$C_{n,3}$	$C_{n,4}$	$C_{n,5}$	-

IV. RESULTS AND ANALISYS

The ECVT sensor scanning process produces capacitance data which is one scanning group by the first 24 electrodes will generate 276 data for conventional scanning method and 552 data for reverse polarity scanning method. The process is repeated up to six times to obtain data as much as 6 × 276 data and 6 × 552 data. The capacitance data is still mixed with the value of stray capacitance whose value varies. In order to be processed and reconstructed into images, the data needs to be normalized using Eq. (4).

$$nc = \frac{C_r - C_l}{C_l} \tag{4}$$

Where nc is the normalized capacitance, C_r is the capacitance to be measured inside sensor, C_l is capacitance at the initial measurement where the sensor has not been filled with the object.

The signal shown in Fig. 7 (a) is the capacitance data of object's by the conventional scanning method of a 64-channel capacitive senor for the first group. The signal pattern is shows the object being scanned appears. The unit of capacitance is expressed in Pico farad. The opposite pair of electrodes will have a low capacitance value, about 0.15pF, while the adjacent

electrode pair has a high capacitance value, about 1.5pF as shown in the graph.

The normalization capacitance by conventional scanning method is shown in Fig. 7 (b) where the data is used in image reconstruction. The X axis is the number of data measurement, whereas the Y axis is the normalized value in which the maximum value is 1. The normalization value represents the value representing the permittivity distribution in the sensor. The smaller value means lower permittivity and vice versa.

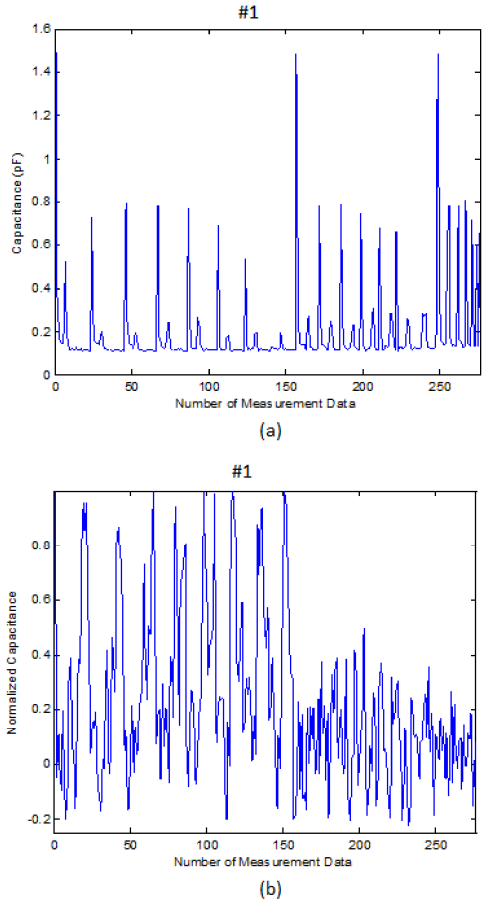


Fig 10. The signal using conventional scanning method from 64-channel capacitive sensor for first group; (a) Capacitance signal; (b) Normalized capacitance

The signal shown in Fig. 8 (a) is a capacitance data by a reverse polarity scanning method from a 64-channel capacitive sensor for the first group. The amount of data is 552 pieces in which the data is twice as much as the conventional method. The normalization graph is shown in Fig. 8 (b) where the data is used in image reconstruction. The X axis is the number of measurement data, whereas the Y-axis is the normalized value whose maximum value is 1.

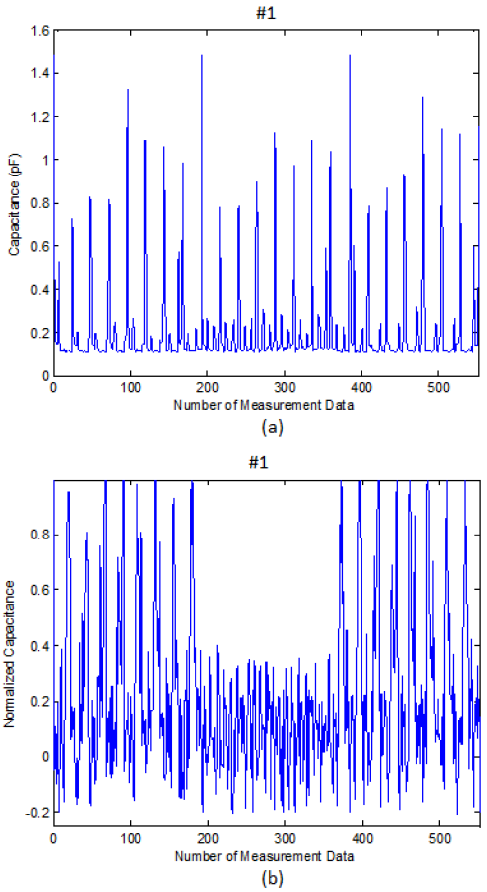


Fig 11. The signal using reverse polarity scanning method from 64-channel capacitive sensor for first group; (a) Capacitance signal; (b) Normalized capacitance

Simulations and experiments have been conducted to determine the performance of the developed ECVT system. The simulation is done using an electromagnetic engineering software to create objects and calculate the matrix sensitivity value. While experiments were performed using an object made of a pvc pipe filled with water (εr = 80). In image reconstruction used ILBP algorithm. To evaluate the performance of the ECVT system, static objects were used to measure parameters such as Mean Absolute Error (MAE), Coefficient-Correlation (R), and image qualitative observation as expressed in Eq. (5) and (6).

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} \quad (5)$$

$$R = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}} \quad (6)$$

Fig. 9 shows the results of image reconstruction for a cylindrical rod object with a combined diameter of 6 cm and 4 cm. Visually shows that the image results by using the reverse polarity scanning method gives better results than using

conventional scanning methods. Each image is divided into three parts consisting of the original image used as a comparator, conventional scan image and reverse polarity image. And also each image consists of two images as three-dimensional images and two-dimensional images in a vertical slicing.

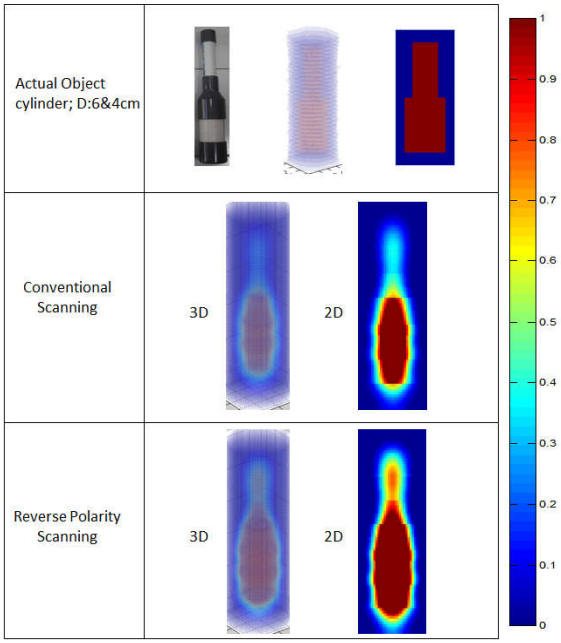


Fig 12. Image reconstruction of cylinder object using conventional and reverse polarity scanning method

Visually shows that the resulting image of all scanned objects is similar to the actual image and also shows that the image produced at the top and bottom ends is slightly convex and pulled up and down. This is due to the characteristic of ILBP algorithm that has elongation error on high contrast dielectrics and if two adjacent objects will have a high error [12]. According to the quantitative measurements, the images generated by reverse polarity scanning method resulting a better image than the conventional scanning method.

TABLE III. QUANTITATIVE PERFORMANCE MEASUREMENT

Scanning Method	Mean Absolute Error (MAE)	Coefficient Correlation (R)
Conventional	0.2281	0.6439
Reverse Polarity	0.2009	0.6645

From the simulation and the experiments can be calculated percentage comparison between the two methods of scanning. The results show that the reverse polarity scanning method gives better reconstruction results with an average correlation coefficient (R) of 3.2% when compared with conventional scanning methods. By reverse polarity scanning method can also reduce error with MAE value of 10.3% when compared with conventional method.

It has been described in the imaging process, image reconstruction is generally done in under sampling measurement i.e. the measurement data is much less than the matrix sensitivity data [13]. This led to the emergence of many artifacts at some point of the sensor. This problem poses a challenge to the development of scanning methods in the ECVT tomography system. Apart from the limitations in the ECVT reconstruction system that still need to be developed, this deficiency promises to be further developed, by improving the design of electrodes and sensors, improving the signal conditioning circuit in order to measure smaller signals, and applying other image reconstruction algorithms such as NN-MOIRT (Neural Network Multi-criteria Optimization Image Reconstruction Technique) [14], [15].

V. CONCLUSION

The design of 64-channel capacitive sensor using new scanning methods on the ECVT tomographic data acquisition system has been tested for qualitative and quantitative performance. The sensor design is useful for multi-phase flow experiments that require long experimental vessel like this design. The experiment is done by comparing conventional scanning methods and reverse polarity scanning method. Simulations and experiments show in general the proposed reverse polarity scanning method provides better reconstructed imagery. From the simulation and experiment result, the average value of Coefficient-Correlation (R) equal to 0.6 and Mean Absolute Error (MAE) is 0.2, so the proposed method can improve the image quality by 3.2%. The advantage of the method is it can increase the amount of measurement data twice than the conventional method so that it can reduce the main problem in the reconstruction algorithm, namely ill-posed. However, this method also increases data acquisition times twice as long as conventional methods. Where conventional scanning methods require acquisition time of 1.52s. Whereas for polarity scanning, exchanging requires acquisition time of 3.03s.

ACKNOWLEDGMENT

The author would like to thank all parties who kindly contribute to this research: Warsito P. Taruno, and Wahyu Widada from CTECH Labs Edwar Technology. This work is supported by a grant from Hibah PITTA DRPM Universitas Indonesia Grant Number: 2362/UN2.R3.1/HKP.05.00/2018.

REFERENCES

[1] D. Sudiana, A. S. Tamsir, and H. Sudibyo, “A digital signal processing algorithm on read out circuit for electrical capacitance tomography,” in *Region 10 Conference (TENCON), 2016 IEEE*, 2016, pp. 1166–1170.

[2] J. M. Weber, K. J. Layfield, D. T. Van Essendelft, and J. S. Mei, “Fluid bed characterization using Electrical

Capacitance Volume Tomography (ECVT), compared to CPFD Software's Barracuda," *Powder Technol.*, vol. 250, pp. 138–146, 2013.

[3] W. Warsito, Q. Marashdeh, and L. S. Fan, "Real Time Volumetric Imaging of Multiphase Flows Using Electrical Capacitance Volume-Tomography (ECVT)," in *Congress on Industrial Process Tomography*, 2007.

[4] A. Wang, Q. Marashdeh, and L. Fan, "ECVT imaging of 3D spiral bubble plume structures in gas□liquid bubble columns," *Can. J. Chem. Eng.*, vol. 92, no. 12, pp. 2078–2087, 2014.

[5] W. P. Taruno *et al.*, "A novel sensor design for breast cancer scanner based on electrical capacitance volume tomography (ECVT)," in *Sensors, 2012 IEEE*, 2012, pp. 1–4.

[6] W. P. Taruno *et al.*, "Brain tumor detection using electrical capacitance volume tomography (ECVT)," in *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*, 2013, pp. 743–746.

[7] W. P. Taruno *et al.*, "Comparisons of Electrical Capacitance Volume Tomography and Ultrasonography for Breast Cancer Detection," *Adv. Sci. Eng. Med.*, vol. 6, no. 8, pp. 845–848, 2014.

[8] Q. M. Marashdeh, F. L. Teixeira, and L.-S. Fan, "Adaptive electrical capacitance volume tomography," *IEEE Sens. J.*, vol. 14, no. 4, pp. 1253–1259, 2014.

[9] Q. M. Marashdeh, Z. Zeeshan, B. Gurlek, and F. L. Teixeira, "AECVT sensors with reconfigurable capabilities for industrial imaging applications," in *Antennas and Propagation (APSURSI), 2016 IEEE International Symposium on*, 2016, pp. 1997–1998.

[10] A. Saputra, W. Widada, and W. P. Taruno, "Rotary sensor system for imaging improvement in electrical capacitance volume tomography," in *Information Technology and Electrical Engineering (ICITEE), 2016 8th International Conference on*, 2016, pp. 1–5.

[11] W. Yang, "Design of electrical capacitance tomography sensors," *Meas. Sci. Technol.*, vol. 21, no. 4, p. 42001, 2010.

[12] W. Warsito, Q. Marashdeh, and L.-S. Fan, "Electrical capacitance volume tomography," *IEEE Sens. J.*, vol. 7, no. 4, pp. 525–535, 2007.

[13] S. Ravishankar and Y. Bresler, "MR image reconstruction from highly undersampled k-space data by dictionary learning," *IEEE Trans. Med. Imaging*, vol. 30, no. 5, pp. 1028–1041, 2011.

[14] Q. Marashdeh, L.-S. Fan, B. Du, and W. Warsito, "Electrical Capacitance Tomography− A Perspective," *Ind. Eng. Chem. Res.*, vol. 47, no. 10, pp. 3708–3719, 2008.

[15] L.-S. Fan, W. Warsito, and B. Du, "Electrical capacitance tomography imaging of gas-solid and gas-liquid-solid fluidized bed systems," *J. Vis.*, vol. 7, no. 1, p. 5, 2004.

Comparison of Cascade and Feedforward-Feedback Controllers for Temperature Control on Stirred Tank Heater Systems

Bhakti Yudho Suprapto
Departement of Electrical Engineering
Universitas Sriwijaya
Jl Raya Palembang-Prabumulih Km. 32 Inderalaya, South of Sumatera, Indonesia
bhakti@ft.unsri.ac.id

Ike Bayusari
Departement of Electrical Engineering
Universitas Sriwijaya
Jl Raya Palembang-Prabumulih Km. 32 Inderalaya, South of Sumatera, Indonesia
ikebayusari@yahoo.co.id

Caroline
Departement of Electrical Engineering
Universitas Sriwijaya
Jl Raya Palembang-Prabumulih Km. 32 Inderalaya, South of Sumatera, Indonesia
caroline.herry@yahoo.com

Muhammad
Departement of Electrical Engineering
Universitas Sriwijaya
Jl Raya Palembang-Prabumulih Km. 32 Inderalaya, South of Sumatera, Indonesia
fadilassegaf@hotmail.com

Abstract*— The temperature control in stirred tank heater (STH) systems is important, especially since the equipment is utilized in a wide variety of industries such as oil and gas. However, within the current STH control process there is often a small disturbance which adversely affects the overall process and it is quite difficult to solve with a regular feedback configuration. Many researchers have applied cascade and feedforward control configurations to solve this problem. Therefore, this paper discusses a comparison of cascade and feedforward-feedback control configurations in maintaining and regulating temperature in the STH. The results show that a cascade controller configuration performs better in various test conditions than a feedforward-feedback controller.

Keywords—cascade; control; feedforward-feedback; PID; stirred tank heater

I. INTRODUCTION

The stirred tank heater (STH) is one of the most widely used types of equipment in the industrial today. The chemical, oil, and gas industries are particularly dependent on STH technology. At the most basic level, the STH is a tank that mixes materials and applies heat to produce a new product. The use of heat is key to the reactions taking place within the STH. Therefore, temperature control becomes an important factor and must be considered when studying ways to produce a good product or otherwise improve the process. Temperature control is usually achieved by controlling a valve which releases fuel; opening the fuel release valve results in higher temperatures while closing the valve has the opposite effect. Temperature control can also be regulated by controlling the flow of heated materials such as gas or steam into or around the STH. Temperature control for STH equipment has been extensively studied including using Proportional Integral Derivative (PID)[1], Proportional Integral (PI)[2], dan fuzzy PID[3] controllers. Other studies reported using a combination of computational intelligent methods to optimize PID control parameters [4]. This controller has been able to solve general problems of temperature control. However, in the process of controlling the temperature, there is often a coincide change in pressure of gas or steam. This change in pressure will interfere with the attempts to control temperature, although the effect does not directly interfere with the system. As a result of this pressure change, the temperature will also change. To solve this problem, the controller configuration is modified by adding a secondary controller which compensates for the initial pressure changes then maintain the correct temperature. In this configuration, the addition of a secondary controller system will also add a secondary sensor to determine the disturbance condition faster than the controlled variable. This approach is called the cascade control system [5]. The advantages of this cascade controller configuration, i.e. the ability to increase the response speed of the primary controller by increasing the response from the secondary controller. If there is a disturbance then the secondary controller will overcome the interference directly.[5][6].

Other disturbances that need to be compensated for during the process within the STH include an error in the modeling and the existence of unmeasurable disturbance. To overcome any modeling error a feedback controller can be used, while to overcome the immeasurable disturbance on the sensor a feedforward controller can be used. Feedforward and feedback controllers can be combined in different ways. In the controller configuration, the feedforward and feedback outputs are summed, where the summed signal will be sent to the controller. Such control is commonly called the feedforward-feedback system[5]. The advantages of the feedforward-feedback controller configuration are: the controlled variable does not have to be measurable; the error correction is immediate when there is a disturbance, and any actions and changes in this configuration do not affect the stability of the process[5][7][8].

Both controller configurations also have the disadvantage i.e. adding to the cost of the equipment, due to the addition of controllers and sensors. However, this addition is proportional to the performance shown by these two control configurations in overcoming the disturbance. This paper compares both types of control configurations in STH temperature control. Both types of control configurations will be seen their response of

control to handle undesirable temperature changes as well as tracking disturbance.

This paper is structured as follows: Section 1—Introduction, explaining the problem, current research, and objectives of this paper; Section 2—Stirred Tank Heater modeling; Section 3—Research Method, describing the controller configuration method used in this paper; Section 4—Results, including an Analysis of the experiment that has been done; and Section 5—Conclusion.

II. Stirred Tank Heater Modelling

STH is a common type of industrial equipment that is designed to mix certain materials at a specific temperature. The contents of the STH must be stirred so that the mixture is evenly distributed in order to consistently produce a particular, high-quality product. In addition to blending the ingredients as they mix, the stirrer also serves to prevent the mixed material from being frozen or becoming too thick. Heat is obtained from heaters, or from steam or gas. The heat conductively moves to the substances contained in the STH tank through the coil wall. The STH work process can be seen in Fig. 1, where temperature (T_i) and flow (F_i) at the input are assumed to be constant. The output is the new product, which is also at a specific flow and temperature. The temperature of the tank is maintained at a setpoint. Heat (Q) is used to keep the temperature constant. Temperature control is performed to regulate the heat by adjusting the valve controlling the flow of steam or gas.

Based on Fig. 1 it is assumed that the heater has been worked and the liquid temperature is kept constant (T_S). The fluid volume must also be maintained at a constant state, a value of V. The mathematical equation model pertaining to the STH can be defined using the law of total energy balance in the process tank by the equation:

$$0 = F\rho C\,(T_{I,S} - T_S) + Q_S \quad (1)$$

Or

$$0 = WC\,(T_{I,S} - T_S) + Q_S \quad (2)$$

Where C is the heat of type (KJ/Kg°C), T_S is the steady state of the output temperature (°C), T_{IS} is the steady state of the input temperature (°C), ρ is the liquid type density (Kg/m^3), F is the fluid flow rate (m^3/s), and W is the mass flow rate (W=F. ρ) (Kg/s).

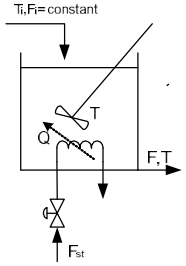


Fig. 1. Stirred Tank Heater Work Diagram

If T_I suddenly increases and if Q_S (energy in steady state condition) has no change, then the energy balance around the tank is::[4][5]

$$V_\rho C\frac{dT}{dt} = F_\rho C(T_1 - T) + Q \quad (3)$$

$$MC\frac{dT}{dt} = WC(T_1 - T) + Q \quad (4)$$

Where M = V. ρ and M are the tank mass (Kg), ρ is the density of the species (Kg / m^3), and V is the volume of the vessel (m^3). The design of this heating system uses the same working principle as a steam-heat modeling process, where input pressure is controlled. In this modeling process, the heater uses gas, which is sent to the stove to burn. So the main factor on the heater is the gas pressure output. The pressure of the gas is then set as the temperature of fire (T_A) through the adjustment of the thermodynamic relationship.

$$T_A = f(P_G) \quad (5)$$

So the energy balance is [4]

$$MC\frac{dT}{dt} = WC(T_I - T) + h_P A_P(T_D - T) \quad (6)$$

$$M_D C_D\frac{dT_D}{dt} = K_A A_A\frac{(T_A - T_D)}{X} - h_P A_P(T_D - T) \quad (7)$$

Where:

- K_A is the Aluminum Conductivity (W / m°C)
- C_D is the Hearth type of wall (KJ / Kg°C)
- $A_P = A_A$ is the cross-sectional area (m2),
- T_A is the temperature in gas (Fire) (°C),
- h_P is the heat transfer coefficient W / m2°C), MD is wall mass (Kg),
- T_D is Tank wall temperature (°C),
- X is reactor tank thickness (m).

Based on the law of energy equilibrium, which stated that the energy change in a system is equal to the amount of energy entering the system minus the amount of energy going out of the system, then equation (8), as follows[4]

$$\frac{d[\rho AhC(T - T_{ref})]}{dt} = \rho FC(T_I - T_{ref}) - \rho FC(T - T_{ref}) + Q \quad (8)$$

or

$$\frac{d[MC\,(T - T_{ref})]}{dt} = WC(T_I - T_{ref}) - WC(T - T_{ref}) + Q \quad (9)$$

The mathematical equation (3) can be developed from the dynamic model equations of the STH heating system, i.e. (8) and (9)[4].

$$MC\frac{dT}{dt} = WC(T_I - T) + h_P A_P\,(T_D - T) \quad (10)$$

$$M_D C_D\frac{dT_D}{dt} = K_A A_A\frac{T_A - T_D}{X} - h_P A_P(T_D - T) \quad (11)$$

Assuming the dynamic conditions in equation (11) can be ignored, it is assumed that the tank wall value of temperature or temperature (T_D) equals the temperature or temperature of the burning gas (T_A). So equation (11) becomes: [4]

$$T_D = \frac{\frac{K_A A_A}{X}T_A + h_P A_p T}{\frac{K_A A_A}{X} + h_p A_P} \quad (12)$$

Substituting into equation (3) will produce,

$$V\rho C\frac{dT}{dt} = WC(T_I) - WC(T) + UA\,(T_A - T) \quad (13)$$

Since the T output variables, UA is a multiplication of heat transfer coefficient (h) to the cross-sectional area (A), and the T_I and P_G input variables are still nonlinear, the variables are converted into a linear form by deviation of variables, resulting in [4]

$$WCT_{I,S} = f(WCT_{I,S}) + WC(T_I\text{-}T_{I,S}) \quad (14)$$

$$UA(P_G - T) = f(UA(P_{G,S} - T_S)) + UA(P_G - P_{G,S}) - (T\text{-}T_S) \quad (15)$$

$$-WCT = -f(WCTS) - WC(T - T_S) \quad (16)$$

The steady state value is

$$0 = WCT_{I,S} - WCT_S + UA(P_{G,S} - T_S) \quad (17)$$

After the linearization values are known, the linear equations (14), (15), and (16) are fed into (13). Substitution with (17) then obtains the deviation equation of the variable below.[4]

$$V\rho C\frac{d(T-T_S)}{dt} = WC(T_I - T_{I,S}) - WC(T\text{-}T_S) + UA((P_G - P_{G,S}) - (T\text{-}T_S)) \quad (18)$$

or

$$V\rho C\frac{dT'}{dt} = WCT_I' - WCT' + UA(P_G' - T') \quad (19)$$

By doing a Laplace transform, generates:

$$V\rho C\, sT'(s) = WCT_I'(s) - WCT'(s) + UA(P_G'(s) - T'(s)) \quad (20)$$

Changing into the transfer function obtains:[4]

$$T'(s) = \frac{K_I}{\tau s+1}T_I'(s) + \frac{K_C}{\tau S+1}\,P_S'(s) \quad (21)$$

The transfer function of the control valve can be seen in the following equation:[4]

$$\frac{P_g(s)}{P(s)} = \frac{K_{CV}}{\tau_{cv}s+1} \quad (22)$$

Where K_{CV} represents the total gain of the control valve, cv denotes the constant time of the control valve (s), P_S represents the output signal from the controller (mA), P_{GS} represents the flow of LPG gas flowing through the valve (psi).

Based on the transfer function equation, the temperature control block diagram of the STH is seen in Fig.2. where the actuator is a control valve that will allow gas or steam to flow in the STH model. The setpoint is the standard temperature to be maintained and kept constant. The stirring speed of the stirrer's driving motor is also made constant. This control is a closed loop. However, an issue arises if there is any disturbance in the pressure of gas or steam, which directly affects the temperature. Resolving this required proper controller configuration.

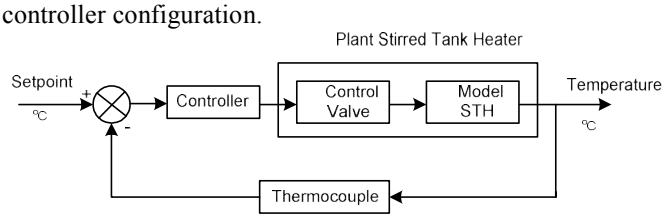


Fig. 2. Block Diagram of Temperatur Control of The Stirred Tank Heater

III. Controller Configuration

The controller which was used in this paper is the Proportional Integral Derivative (PID) controller because it is reliable, simple, and widely used throughout various industries. However, the configuration will compare the performances between cascade and feedforward.

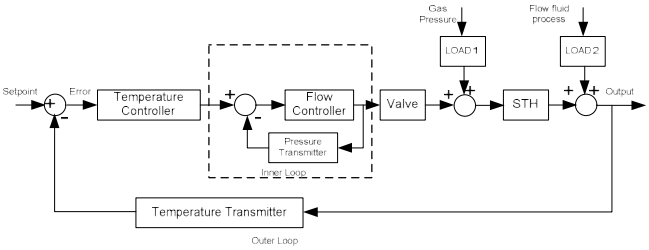


Fig. 3. Block Diagram of Cascade Control

A. Cascade

The block diagram of the controller with this cascade configuration can be seen in Fig. 3. In this controller, there are two feedback paths in the cascade control system, thus forming two loop controls, namely the outer loop which is the primary loop (or master), and the inner loop which is the secondary loop (or slave). The master or primary loop controls the primary variable process (fluid temperature process), while the slave or secondary loop controls the process of secondary variables (vapor or gas pressure).

The cascade controller is chosen because there are several reasons such as the output response of the single control is not as expected, there is the addition of a secondary variable in the control of the plant, which is added as a plant control, can overcome the disturbance.

B. Feedforward-feedback

In this STH control system, if there is change of load from the plant while the flow of steam or gas is constant, so it will be used a feedback control system. But if the load changes happen too quickly, then another type of control system becomes necessary. Because conventional feedback paths need to "see" errors before performing corrective actions, a standard feedback system will not be able to handle the load well if the load change frequency is too fast. Therefore, the design of the feedforward system as shown in Fig. 4 is based on process requirements, reducing the large loads that can disrupt the output process. The ultimate design of the system is left up to the designer because it is expected that a holistic, streamlined design will be able to reduce the interaction between feedforward and feedback (FFFB).

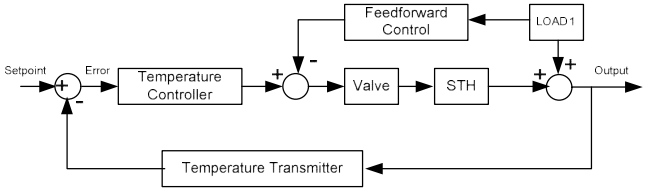


Fig. 4. Block Diagram of the Feedforward-feedback Control

IV. Experiment Result

The purpose of this paper is to compare the configurations of the cascade and feedforward-feedback controllers. Both of these configurations use PID controllers as both primary and secondary controllers. The cascade controller has a response when given setpoint temperature of 90°C. As shown in Fig. 5, the resulting overshoot value is 12.1%, the rise time is 1.075 seconds. The settling time value is 25.45 seconds, and the steady state error is 0% and the resulting offset value is 0.

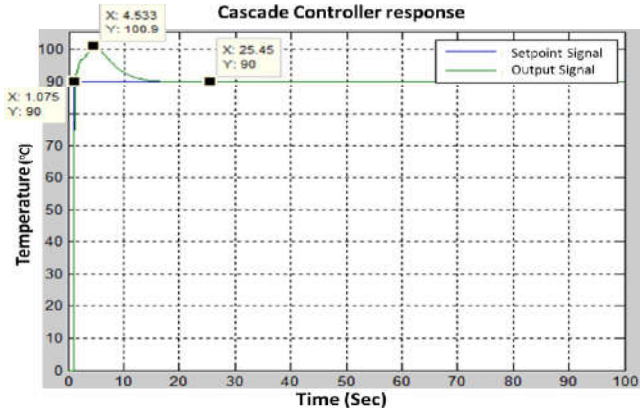


Fig. 5. The response of Cascade Control with setpoint

Then it tested the disturbance in the form of a changing setpoint as seen in Fig. 6. This test obtained the results of overshoot values of 12.1%, 4.64%, 3.83%, 3.24% and 0.2%. The average overshoot generated is 4.8%. The time rise results are 0.2836 seconds, 0.28 seconds, 0.3 seconds, 0.3 seconds and 0.3 seconds. This results in an average rise time of 0.292 seconds. The settling time value results are 24.4 seconds, 21.73 seconds, 22.1 seconds, 21.1 seconds and 20.6 seconds. This yields an average required settling time of 21.98 seconds. The steady state error is 0% and the offset value is 0. These results indicate the cascade control system on the stirred tank heater can work well when used with changing set-point conditions. This is evidenced by low overshoot value, fast rise time and settling time and 0% steady state error. This proves an STH will work well under these conditions.

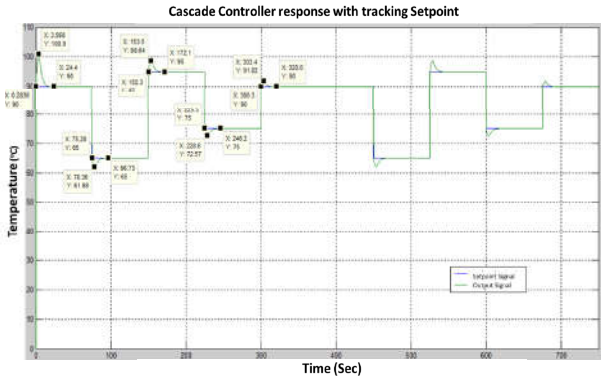


Fig. 6. The response of Temperature Control of STH using Cascade Control with Tracking Set-point

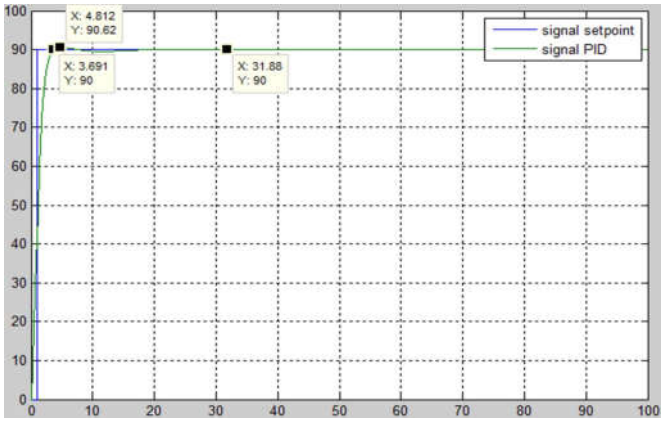


Fig. 7. The response of Temperature Control of STH using FFFB with setpoint

Then, comparison tests were run using the FFFB controller, again with both a single setpoint and a tracking setpoint. Based on the output response generated in Fig 7, it can be seen that the FFFB controller was effective for single-setpoint conditions. The resulting overshoot value is 0.68%, the rise time is 3.691 seconds, the settling time is 31.88 seconds, the steady state error is 0% and the offset value result is 0. This response is very good, the overshoot is very small, almost 0%, the rise time is fast, the settling time is also quite good and the steady state error is 0%. This proves that the FFFB control system is good enough to overcome the existing issues.

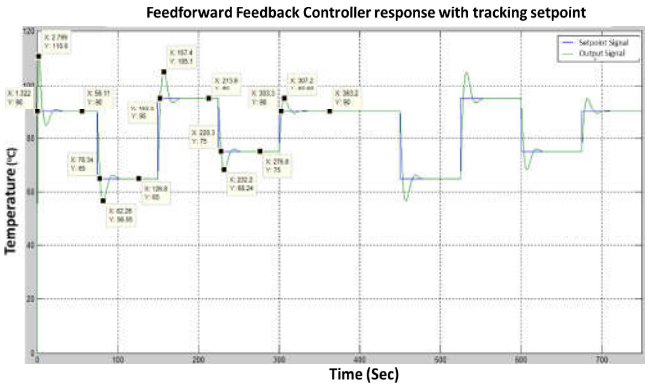


Fig. 8. The response of Temperature Control of STH using FFFB with Tracking Set-point

However, in Fig.8, the results of the FFFB configuration with a changing setpoint were not as promising. The resulting overshoot value is 22.8%, 13%, 10.6%, 9.01% and 5.6%, so the average overshoot generated is 12.2%. The time rise results were 1.322 seconds, 3.34 seconds, 3.3 seconds, 3.3 seconds and 3.3 seconds, which yields an average rise time equal to 2.91 seconds. Settling time values were 56.11 seconds, 51.8 seconds, 63.6 seconds, 51.8 seconds and 63.2 seconds, so the average settling time required is 57.3 seconds. The steady state error is 0% and the resulting offset value is 0. The resulting response is good enough, the overshoot is low, the rise time is fast and the steady state error is 0%, but the settling time is not very fast since the average was 39.81 seconds.

Based on the response generated by these two configurations, it can be seen that for conditions where the temperature setpoint is desired to be kept constant, the cascade controller has a smaller rise time and faster settling time than the FFFB controller. Therefore, the cascade is better than FFFB even though the overshoot generated is quite high. The comparison of response both of this controller can be seen in table 1. Although better, the cascade has its drawbacks; with its high overshoot, equipment may be adversely affected. In the cascade configuration, responses are accelerated due to the direct influence of the secondary controller. Anticipating the influence of pressure changes on steam or gas in the heater results in a high overshoot in an attempt to reach the new setpoint temperature. In the FFFB controller tests, under fixed setpoint conditions, the results indicate that the controller tries simply to approach the setpoint. Therefore, adding FFFB controllers is quite effective in minimizing overshoot. This ensures the safety of the equipment, but the time to reach a stable temperature exceeds the amount of time needed by the cascade. Similarly, when comparing the results of the test with changing setpoints, the cascade configuration proved to be the better choice because the rise time, overshoot and settling time all were lower than the FFFB. The only advantage to FFFB in the conditions introducing disorder by testing setpoint changes is that FFFB can eliminate the overshoot and thereby reduce the risk of equipment damage compared to the cascade.

TABLE I. COMPARISON OF CASCADE AND FEEDFORWARD-FEEDBACK CONTROLLER RESPONSE WITH SETPOINT

Controller	Rise time (second)	Overshoot (%)	Settling time (second)	Error Steady state (%)
Cascade with setpoint	1.575	12.1	25.45	0
FFFB with Setpoint	3.691	0.68	31.88	0

V. CONCLUSIONS

Based on the experiments and the tests that have been conducted, the cascade is better than the feedforward-feedback control system for a stirred tank heater (STH). The biggest advantage is that the cascade system generates the lowest rise time and settling time values. The cascade also has better results than the feedforward-feedback under changing setpoint conditions, based on the following factors: overshoot situations, rise time, and lower settling time. Future jobs will test other types of controllers such as PI and fuzzy logic, to develop more and better controllers for STH systems.

REFERENCES

[1] A. S. Rajagopalen, "Identification of an effective controller for a stirred tank heater process," *Int. J. Eng. Adv. Technol.*, vol. 3, no. 1, pp. 271–279, 2013.

[2] S. Sehgal and V. Acharya, "Design of PI controller for continuous stirred tank heater process," in *2014 IEEE Students' Conference on Electrical, Electronics, and Computer Science, SCEECS 2014*, 2014.

[3] K. V and G. D. G, "Comparative Analysis of PID and Fuzzy PID Controller Performance for Continuous Stirred Tank Heater," *Indian J. Sci. Technol.*, vol. 8, no. 23, pp. 1–7, 2015.

[4] N. I. Septiani, I. Bayusari, C. Caroline, T. Haiyunnisa, and B. Y. Suprapto, "Optimization of PID control parameters with genetic algorithm plus fuzzy logic in the stirred tank heater temperature control process," in *ICECOS 2017 - Proceeding of 2017 International Conference on Electrical Engineering and Computer Science*, 2017, pp. 61–66.

[5] D. E. Seborg, T. F. Edgar, D. A. Mellichamp, and F. J. Doyle III, *Process Dynamics, and Control*. 2011.

[6] I. Kaya, N. Tan, and D. P. Atherton, "Improved cascade control structure for enhanced performance," *J. Process Control*, vol. 17, no. 1, pp. 3–16, 2007.

[7] M. Malek, P. Makys, and M. Stulrajter, "Feedforward Control of Electrical Drives - Rules and Limits.," *Power Eng. Electr. Eng.*, vol. 9, pp. 35–42, 2011.

[8] J. L. Guzmán, T. Hägglund, M. Veronesi, and A. Visioli, "Performance indices for feedforward control," *J. Process Control*, vol. 26, pp. 26–34, 2015.

Detecting Business Process Anomaly Using Graph Similarity Based on Dice Coefficient, Vertex Ranking and Spearman Method

Cecilia Stephanie
Department of Information Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
ciliata7@gmail.com

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

***Abstract*—Graph similarity is a method to check the graphic equations between two graphs. Anomaly detection in a business process can use the graph similarity method as well. The graph of the ongoing business process compared to the Standard Operating Procedure's graph (SOP's graph) as a reference. Standard Operating Procedures (SOPs) can definitely as a benchmark of a company running a business process. In this study, the case taken to detect anomaly of business processes in the process of replacing parts in wireless device distributors. The method used is dice coefficient to detect vertex or edge overlap and vertex ranking using the Spearman method to detect missing or altered vertices. The result of detecting business process anomaly using dice coefficient found three business process anomalies. While in the vertex ranking method using Spearman coefficient found two anomalies. Anomalies were found in vertex ranking method are the same as the expert judgement's results.**

Keywords—graph similarity;business process anomaly;dice coefficient;Vertex Ranking, Spearman method

I. INTRODUCTION

Business processes that exist in the company must compared with the applicable Standard Operating Procedure (SOP). SOP serves as a guide how business processes should be working. In fact, the business processes that run in the field are not always the same as the existing SOP. Deviation distance between SOP and prevailing business processes called anomaly. A deviation is definitely an anomaly when it exceeds the firm-specified thresholds. Furthermore, anomaly also used as a guide to update SOP or even find fraud from the current business process.

One way that used to detect anomaly is by the graph similarity method. Graph similarity is a method by comparing the similarities of SOP graphs and business processes. The SOP graph consists of vertices and edges. Vertex describes every task that acts as a complex object connected by edges. Edge is the interaction between two objects or vertices [1]. Graph similarity used to detect missing vertices, missing edge, vertices modification and edges modification [2].

The deviation of SOP graphics and business process can extracted with Vertex / Edge Overlap algorithm and Vertex Ranking algorithm. Vertex / Edge Overlap calculates the distance between SOP graph and business process graph with the Dice Coefficient method. On the other hand, Vertex Ranking calculates the distance between graphs with Spearman method. After calculated, the deviations of the three methods above compared. Before drawing a conclusion, it is necessary to establish a similarity threshold that will separate the anomaly and the graph that considered almost similar [3]. Determination of threshold in this study using Process-Based Fraud (PBF) rating. The results of the experiments will compared with the expert judgment of the company.

II. LITERATURE

A. Standard Operating Procedure (SOP)

Standard Operating Procedure (SOP) can define as "a set of written and detailed instructions that document a routine or repetitive activity followed by an organization to achieve uniformity of the performance of a specific function" [4]. SOP ensures all workers perform tasks in the same way and the process must produce consistent output. If it does not run correctly, then the whole operation will fail [5]. SOPs allow organizations to be consistent and consistent with cross-departmental processes that will impact on products and services. Unlike before, Edelson & Bennett argued that SOPs were a management effort to improve the results of the process consistently and efficiently [5]. If there is no SOP, then there is no measure for the company to determine whether the running process is the best work process. Therefore, SOP must be reviewed periodically whether it is still valid at this time or not.

B. Graph Similarity

Graph similarity is applied to many fields and is measured by various similarity measurement [6]. One of the areas that apply similarity is a business process model. The graph similarity technique can divided into three categories with distance, iterative and feature extraction. In this paper, we use it for determine distance between SOP graph and business process graphs. For example, we have two graphs that have similarity checked. These two graphs have the possibility of having different vertices and edges. The graph similarity approach will produce similarity measurement values between "0" and "1" [7]. The graph differs or suspected as anomaly if

value is close or equal to "0". Conversely, if the value is close or equal to "1" can be defined the same graph.

C. Anomaly

Anomalies can definitely use patterns that do not conform to normal behavior [8]. There are several factors from [8], which make the anomaly difficult to detect, such as:

- The boundary between the anomalous case and the normal case is not too obvious.
- Normal state influenced by many factors causing the change of anomaly boundary in the future.
- Data confounding is difficult to identify and causes anomaly indications to become chaotic.

That challenges makes anomaly detection difficult. Most existing anomaly detection techniques identify anomalies based on the nature of the data, availability of label data and anomaly types to be detected. In this paper, we try to detect anomalies by setting threshold according to PBF rating reference. Furthermore, the anomalous results of the graph similarity will compared with the expert judgment of the company.

D. PBF rating

Detecting process-based fraud (PBF) can see from three different angles, there are the point of view of the business process, business role and organization [9]. Comparing different business processes with respective models is the first one. Second, PBF analyzes the deviation of the business role. Third, PBF analyzes the people who do the tasks according to the division of tasks. The superiority of PBF can compare business processes and SOPs [10]. PBF can also detect missing vertices or edges. In this paper, we only use the first approach to check the process (the point of view business process). The threshold can be determined from the score of risk. The risk scores obtained from the number of business processes detected, divided by a hundred then multiplied by the weight factor [11]. Weight factor usually set by expert judgment in company because that person knows the risk factors in their business processes. There are three approaches to determine the weight factor is pressure, opportunity, and rationalization [12].

III. Proposed Method

A. Vertex / Edge Overlap

The vertex / edge overlap method approach is to calculate the number of vertices or edges of two graphs [2]. The advantage of this method is that it can detect the missing vertices or edges of the graphs tested with SOP graphs created as a comparison. It is undeniable that there are deficiencies in the vertex / edge overlap method of looking for ignore the sensitivity and coverage requirements. Similarity can be measured using Jaccard index or Dice coefficient. The result range of Jaccard and Dice method is equally between 0-1. The difference between Jaccard index and Dice coefficient is on the dice method of comparing the intersection to the total number tested. While the Jaccard index, compare the intersection with the disjunction of the data tested. The intersection of the graph is the data that is on both graphs. Disjunction is a non-slice diagram. Union is a composite data of two graphs or more precisely the sum of intersection and disjunction. The difference of intersection, union, and disjunction described in Figure 1.

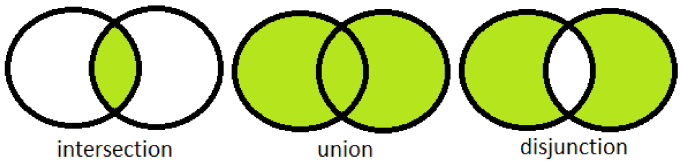


Figure 1. The difference of intersection, union, and disjunction

In this paper, we choose to use Dice coefficient because it's more to describe the similarity of graph than the Jaccard index because there may be value outside disjunction. The formula used is twice the sum of the vertices and edge intersection and then divided the total vertices and edges of the two graphs. The formula Vertex/Edge Overlap between SOP's graph (G) and business process case's graph (G')

$$Dice\,(G, G') = 2\,\frac{(V \cap V') + (E \cap E')}{(V + V') + (E + E')} \qquad (1)$$

Where:

V = total vertices in graph G
V' = total vertices in graph G'
E = total edges in graph G
E' = total edges in graph G'

B. Vertex Ranking

The definition of vertex ranking is a method of comparing vertices of two graphs by ranking. Vertex on both graphs will rank based on whether or not there is a vertex in the business process. In paper [2], Spearman method usually applied to connect a ranking variable. The use of such methods is still rare for graph similarity. First, vertex will sort based on the priority ranking of vertices that must exist. Then, look for the distance between the SOP's vertex and the business process case's vertex. Distance from the vertices of SOP and business process vertices squared and entered into superman equations. Spearman method formula comparing SOP's graph and business process cases described in

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)} \qquad (2)$$

Where:

r_s = Spearman coefficient
d_i^2 = square distance between SOP's vertex and business process case's vertex
n = total vertex that has the same rank

In the above equation, distance symbolized by di, while n is the number of vertices in the graph. Equation in (2) only be used if there is no same ranking value on one graph, if there is the same ranking value in the graph use formula

$$\eta = \frac{\sum x^2 + \sum y^2 - \sum di^2}{2\sqrt{\sum x^2 . \sum y^2}} \quad (3)$$

Where:

x^2 = square of total vertices in the SOP graph

y^2 = square of total vertices in the business process case graph

di^2 = square distance between SOP's vertex and business process case's vertex

The x^2 in equation obtained from n squared three, then reduced by n, detail equations can be look in (4) and (5).

$$\sum x^2 = \frac{n^3 - n}{12} - \sum Tx \quad (4)$$

$$\sum y^2 = \frac{n^3 - n}{12} - \sum Ty \quad (5)$$

Where:

x^2 = square of total vertices in the SOP graph

y^2 = square of total vertices in the business process case graph

n = total vertex that has the same rank

$\sum Tx$ = total of the average SOP's vertices twin rank

$\sum Ty$ = total of the average business process case's vertices twin rank

Then, divided it by twelve then reduced by the average twin ranking value ($\sum Tx$). In equation (6), $\sum Tx$ is defined by the number of twin rankings for each ranking (t), it squared by three and reduce with (t), last the result divided by twelve.

$$Tx / Ty = \frac{t^3 - t}{12} \quad (6)$$

Where:

Tx = the average SOP's vertices twin rank

Ty = the average business process case's vertices twin rank

t = total vertices that have twin rank

Vertices that do not exist on other graph would be easy to detect. The challenge of this method is to rank appropriate for business processes.

IV. Experiment and results

Business process taken for this experiment is the replacement parts of wireless devices on the distributor of wireless devices. The study will compare SOPs with 10 cases of running business processes. In SOP, there are 23 tasks with 28 vertex and 58 edges. Task of SOP's replacement parts of wireless devices described in Table 1. SOP's graph illustrated in Figure 2.

Table 1. Task of SOP's replacement parts of wireless devices

Task	Activity
t1	Request parts replacement
t2	Check device
t3	Inform status device
t4	Create device receipt
t5	Fill the parts replacement registration
t6	Deliver device to the technician
t7	Service device
t8	Check service results
t9	Calculate cost
t10	Make a service offer
t11	Run billing SOP
t12	Create delivery order
t13	Check the warranty
t14	Make notification of damaged device
t15	Inform replacement device to the customer
t16	Request device replacement to vendor
t17	Replacement parts
t18	Check replacement stock of parts
t19	Ask for approval replacement
t20	Fill in the parts request form
t21	Check sales stock of parts
t22	Request stock release
t23	Create job costing

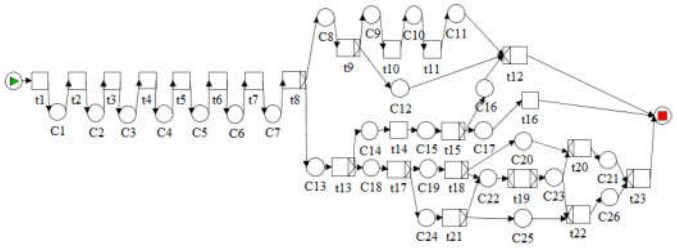


Figure 2. Graph for SOP's replacement parts of wireless devices

The SOP graph illustrates all possible business processes at the company. This means not all tasks must missed. One case study of anomaly detection is Figure 3. The first step of anomaly detection Vertex / Edge Overlap by Dice coefficient method is calculate the number of vertices and edges of the SOP graph (Figure 2). The second step is to calculate the vertex and edge intersections of the SOP charts and business processes. Finally, enter the results of the first and second steps into the formula (1). Example for case study 1 can see in Figure 3.

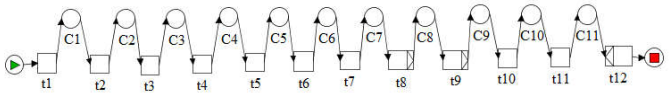


Figure 3. Case study 1

Before the experiment, we must set the threshold which one business process stated anomaly. Anomalies detect by the risk factor. Each risk factor in business process given weight score. In one hundred experiment, ten cases determined as anomaly which weight factor for that risk factor is five. Then, we get 0.5 for threshold. When the result of an experiment with dice coefficient methods or vertex ranking using Spearman method greater or equal to 0.5, we identify similar business processes, and less than 0.5 for anomaly.

Dice coefficient used to find overlapping vertices and edges. The number of vertices in the SOP is 28 vertices, while the number of edges in the SOP is 58 edges. The number of vertices and edges in the case study explained in Table 2. Column $V \cap V'$ shows the number of vertices overlap between SOP and each case studies. Column $E \cap E'$ shows the number of edges overlap between SOP and case studies. The higher the value of vertices and edges overlap, then the degree of similarity SOP's graph and business process case study's graph higher. Then the data in Table 2 entered into the equation (1) to get the dice coefficient.

TABLE 2. The number of vertices and edges in the case study

	V	V'	E	E'	$V \cap V'$	$E \cap E'$
SOP vs case 1	28	11	58	24	11	24
SOP vs case 2	28	9	58	20	9	20
SOP vs case 3	28	11	58	24	11	24
SOP vs case 4	28	11	58	24	11	24
SOP vs case 5	28	12	58	26	12	26
SOP vs case 6	28	13	58	29	13	29
SOP vs case 7	28	13	58	29	13	29
SOP vs case 8	28	8	58	18	7	16
SOP vs case 9	28	7	58	16	3	8
SOP vs case 10	28	12	58	26	7	16

The result of graph similarity calculation using the Dice Coefficient method is in the range 0.2 to 0.65. With the determination of the threshold of 0.5, cases in eighth, ninth, and tenth can categorized as anomaly. Once checked, there are many missing vertices and edges, also many modification vertices and edges. Overall, result of low similarity maximum only 0.65. This is because there are many branches in the SOP that only need to be selected one branch.

The first step in the Vertex Ranking method is rank for each task in business process. After that, we can get the distance, which uses in Spearman method. A rank of SOP and each task in that business process case study showed in Table 3. Distance and square distance from each task in business process case to SOP described in Table 4 and Table 5. After getting the square distance, we input the value to the equation (3).

TABLE 3. Rank of SOP and each task in that business process case study

Task	**SOP**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
t1	12	9	7	8	8	10	11	11	7		11
t2	14	11	9	9	9	11	12	12			12
t3	11	8	6	6	6	9	10	10			10
t4	1	1	1	1	1	1	1	1	1		
t5	8	6	4	4	4	7	8	8	4	4	8
t6	10	7	5	5	5	8	9	9	5	5	
t7	13	10	8	10	10	12	13	13	8		
t8	9	5	3						3		
t9	2	2									
t10	3	3									
t11	4	4									
t12	4		2						2	2	
t13	8			5	5	5	7	7			
t14	2			2	2						
t15	7			7	7						
t16	4			3							
t17	3				3						
t18	2					2	2	2			
t19	5					4	4				4
t20	6						5				5
t21	4						6				3
t22	3					3	3				
t23	7					6					
t24	5							5			
t25	6							6			
t26	4							4			
t27	3							3			
t28	7										
t29	0								6	6	
t30	0									7	
t31	0									1	
t32	0									3	
t33	0										1
t34	0										2
t35	0										7
t36	0										9
t37	0										6
#case	28	11	9	11	11	12	13	13	8	7	12

TABLE 4. Distance from each task in business process case to SOP

Task	1	2	3	4	5	6	7	8	9	10
t1	3	5	4	4	2	1	1	5	12	1
t2	3	5	5	5	3	2	2	14	14	2
t3	3	5	5	5	2	1	1	11	11	1
t4	0	0	0	0	0	0	0	0	1	1
t5	2	4	4	4	1	0	0	4	4	0
t6	3	5	5	5	2	1	1	5	5	10
t7	3	5	3	3	1	0	0	5	13	13
t8	4	6	9	9	9	9	9	6	9	9
t9	0	2	2	2	2	2	2	2	2	2
t10	0	3	3	3	3	3	3	3	3	3
t11	0	4	4	4	4	4	4	4	4	4
t12	4	2	4	4	4	4	4	2	2	4
t13	8	8	3	3	3	1	1	8	8	8
t14	2	2	0	0	2	2	2	2	2	2
t15	7	7	0	0	7	7	7	7	7	7
t16	4	4	1	4	4	4	4	4	4	4
t17	3	3	3	0	3	3	3	3	3	3
t18	2	2	2	2	0	0	0	2	2	2
t19	5	5	5	5	1	1	5	5	5	1
t20	6	6	6	6	6	1	6	6	6	1
t21	4	4	4	4	4	-2	4	4	4	1
t22	3	3	3	3	0	0	3	3	3	3
t23	7	7	7	7	1	7	7	7	7	7
t24	5	5	5	5	5	5	0	5	5	5
t25	6	6	6	6	6	6	0	6	6	6
t26	4	4	4	4	4	4	0	4	4	4
t27	3	3	3	3	3	3	0	3	3	3
t28	7	7	7	7	7	7	7	7	7	7
t29	0	0	0	0	0	0	0	22	22	0
t30	0	0	0	0	0	0	0	0	21	0
t31	0	0	0	0	0	0	0	0	27	0
t32	0	0	0	0	0	0	0	0	25	0
t33	0	0	0	0	0	0	0	0	0	27
t34	0	0	0	0	0	0	0	0	0	26
t35	0	0	0	0	0	0	0	0	0	21
t36	0	0	0	0	0	0	0	0	0	19
t37	0	0	0	0	0	0	0	0	0	22

TABLE 5. Square distance (di^2) from each task in business process case to SOP

	1	2	3	4	5	6	7	8	9	10
t1	9	25	16	16	4	1	1	25	144	1
t2	9	25	25	25	9	4	4	196	196	4
t3	9	25	25	25	4	1	1	121	121	1
t4	0	0	0	0	0	0	0	0	1	1
t5	4	16	16	16	1	0	0	16	16	0
t6	9	25	25	25	4	1	1	25	25	100
t7	9	25	9	9	1	0	0	25	169	169
t8	16	36	81	81	81	81	81	36	81	81
t9	0	4	4	4	4	4	4	4	4	4
t10	0	9	9	9	9	9	9	9	9	9
t11	0	16	16	16	16	16	16	16	16	16
t12	16	4	16	16	16	16	16	4	4	16
t13	64	64	9	9	9	1	1	64	64	64
t14	4	4	0	0	4	4	4	4	4	4
t15	49	49	0	0	49	49	49	49	49	49
t16	16	16	1	16	16	16	16	16	16	16
t17	9	9	9	0	9	9	9	9	9	9
t18	4	4	4	4	0	0	0	4	4	4
t19	25	25	25	25	1	1	25	25	25	1
t20	36	36	36	36	36	1	36	36	36	1
t21	16	16	16	16	16	4	16	16	16	1
t22	9	9	9	9	0	0	9	9	9	9
t23	49	49	49	49	1	49	49	49	49	49
t24	25	25	25	25	25	25	0	25	25	25
t25	36	36	36	36	36	36	0	36	36	36
t26	16	16	16	16	16	16	0	16	16	16
t27	9	9	9	9	9	9	0	9	9	9
t28	49	49	49	49	49	49	49	49	49	49
t29	0	0	0	0	0	0	0	484	484	0
t30	0	0	0	0	0	0	0	0	441	0
t31	0	0	0	0	0	0	0	0	729	0
t32	0	0	0	0	0	0	0	0	625	0
t33	0	0	0	0	0	0	0	0	0	729
t34	0	0	0	0	0	0	0	0	0	676
t35	0	0	0	0	0	0	0	0	0	441
t36	0	0	0	0	0	0	0	0	0	361
t37	0	0	0	0	0	0	0	0	0	484
	497	626	535	541	425	402	396	1377	3481	3435

The results of calculations using Spearman method vary greatly between 0.35 to 0.89. The difference is due to ranking first on the vertex, so the determination of missing vertex and vertex modification is more accurate. At the time of ranking, there are some difficulties if a vertex exists only one diagram and if there is the same ranking value. The solution to such difficulties is to rank the most recent SOPs for vertices that exist only in the case of business processes. If there is the same ranking value, then use Spearman method formula on (3). With the determination of the threshold value of 0.5 then obtained cases ninth and tenth have anomaly tendencies.

The results of the Vertex Ranking using Spearman method in this study are very good because it is able to produce data that really contrast. It cannot be denied that the vertex ranking is very dependent on the selection of task necessity exists as a ranking reference. If ranking is not correct mapping, then Spearman method failure rate will be very high. The results of Dice Coefficient and Spearman method calculations shown in Table 6.

TABLE 6. Results of computing graph similarity

Case Study	**Dice Coefficient**	**Vertex Ranking using Spearman method**
1	0.57851	0.86341

2	0.50435	0.82796
3	0.57851	0.85295
4	0.57851	0.85130
5	0.61290	0.88320
6	0.65625	0.88952
7	0.65625	0.89117
8	0.41071	0.65954
9	0.20183	0.35958
10	0.37097	0.42352

In the case study ninth, the lowest dice coefficient is 0.20183. This is because the number of intersections between vertices between SOP and business process case study having the smallest value is three (see in Table 2). The number of intersections between the edges between SOP and business process case study having the smallest value too, is eight. In contrast, in the case studies sixth and seventh had the highest number of vertex overlaps with SOPs, thirteen. The highest number of edge overlaps also in the case studies sixth and seventh, was twenty-nine compared to the SOP.

The vertex ranking value based on the Spearman coefficient has the highest value, 0.89117 in the case study seventh. That is because the ranking of tasks in case study seventh has a rating that is almost the same as SOP's tasks (see in Table 3). The lowest value of the vertex ranking based on the Spearman coefficient is in case study ninth, which is 0.35958. In case ninth, there were four tasks were not in the SOP. Tasks that are not in the SOP's graph ranked as null or zero. This causes the distance value between the ninth business process's graph and the SOP's graph to be high and produces a low value of Spearman coefficient.

V. CONCLUSION

The conclusion from the research that the algorithm that used to compare business process graph is dice coefficient and Spearman method. Dice coefficient is suitable for binary vectors or two sets of data, X and Y. The result of comparing ten cases of business processes with dice coefficient found there are three anomaly processes with threshold 0.5, cases eighth, ninth and tenth.

In other hand, Spearman method obtained by ranking for every vertex in business process and SOP and then get distance from it. Spearman method measurements also found two cases of anomaly business processes (under threshold 0.5), cases ninth and tenth. After comparison with expert judgement, found that only case ninth and tenth were anomalies.

It was concluded that use several methods can obtain more accurate anomaly indicators than one method. In the next study, we can add the method of Hamming distance and Jaccard index.

REFERENCES

[1] Chen, H. H., Gou L., Zhang X.L., and Giles, C.L. "Discovering Missing Links in Networks Using Vertex Similarity Measures". *Proceedings of the Twenty-Seventh Annual ACM Symposium on Applied Computing (SAC'12), Trento* (2012):138–143. **DOI:** 10.1145/2245276.2245305

[2] Papadimitriou, P., Dasdan, and A., Garcia-Molina, H. "Web graph similarity for anomaly detection." *J Internet Serv Appl* 1.1(2008):1167. **DOI:** 10.1007/s13174-010-0003-x

[3] Sarno, R., and Sinaga, F. "Business Process Anomaly Detection using Ontology-based Process Modeling and Multi-level Class Association Rule Learning". *International Conference on Computer, Control, Informatics and its Applications (IC3INA)* 2015: 12-17. **DOI:** 10.1109/IC3INA.2015.7377738

[4] Amare, Gidey. "Reviewing the Values of a Standard Operating Procedure." *Ethiopian Journal of Health Sciences* 22.3 (2012): 205–208. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3511899/

[5] Treville, S.D., Antonakis, J., and Edelson, N.M. "Can Standard Operating Procedures be Motivating? Reconciling Process Variability Issues and Behavioural Outcomes". *Total Quality Management* 16.2 (2005):231–241. **DOI:** 10.1.1.549.8898

[6] Koutra, D., Parikh, A., Ramdas, A., and Xiang, J. "Algorithms for graph similarity and subgraph matching". Technical report. Carnegie-Mellon-University (2011). **DOI:** 10.1.1.377.4291

[7] Broder, A.Z., Glassman, S.C., Manasse, M.S., and Zweig, G. "Syntactic Clustering of the Web". In Proc. of Int. World Wide Web Conf. (WWW), Apr 1997: 393-404. **DOI:** 10.1016/S0169-7552(97)00031-7

[8] Sensarma, D., and Sarma, S.S. "A Survey on Different Graph Based Anomaly Detection Techniques". *Indian Journal of Science and Technology* 8.31 (2015). **DOI:** 10.17485/ijst/2015/v8i31/75197

[9] Huda, S., Sarno, R., and Ahmad, T. "Fuzzy MADM Approach for Rating of Process-Based Fraud". *Journal of ICT Research and Applications* 9.2 (2015): 111-128. **DOI:** 10.5614/itbj.ict.res.appl.2015.9.2.1

[10] Huda, S., Sarno, R., and Ahmad, T. "Increasing Accuracy of Process-based Fraud Detection Using a Behavior Model". *International Journal of Software Engineering and Its Applications* 10.5 (2016): 175-188. **DOI:** 10.14257/ijseia.2016.10.5.16.

[11] Rahmawati, D., Yaqin, M.A., Sarno, R.. "Fraud Detection on Event Logs of Goods and Services Procurement Business Process Using Heuristics Miner Algorithm". *International Conference on Information & Communication Technology and Systems (ICTS)* 2016: 249 – 254. **DOI:** 10.1109/ICTS.2016.7910307

[12] Rahmawati, D., Sarno, R., Fatichah, C., Sunaryono, D. "Fraud Detection on Event Log of Bank Financial Credit Business Process using Hidden Markov Model Algorithm". *International Conference on Science in Information Technology (ICSITech)* 2017: 35-40. **DOI:** 10.1109/ICSITech.2017.8257082

Job Performance Analysis Using Factor Evaluation System and Process Mining

Gabriel Sophia
Department of Management Technology
Faculty of Bussiness and Management Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
Sophia27gabriel@gmail.com

Riyanarto Sarno
Department of Informatic
Faculty of Information and Communication Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

***Abstract*— Job performance is a method to evaluate the performance such as a company or employee by comparing the work that has been finished in that several period with the standard of company. Factor Evaluation System method is one of popular method to used because it has 9 factors to analyze a position. But FES no more detail in spell out each factor, so in this research using process mining to enrich FES application. Process mining using event log that has history of case, activity, time stamp and resources of employee. This paper using AHP to get weight for every FES factor, so each factor have standard to be achieve. Considering the advantages of FES and process mining, this paper proposes the combination FES and process mining to more detailing the factors of FES, which are required for recruitment using factor all factor that have weight; employee monitoring using factor supervisory controls(2), guidelines(3) and personal contacts(6); for employee training using factor knowledge required by the position(1), personal contacts(6), purpose of contacts(7) and work environment(9); remuneration using factor complexity(4), scope and effect(5) and physical demands(8).**

Keywords*— *Job Performance, Factor evaluation System, Process Mining

I. INTRODUCTION

Job performance is one of the most important dependent variables and has been studies for a long decade [1, 2, 3, 4]. There are two types behavior of employee that required to identified job performances for Organizational effectiveness : contextual performance and task performance. Task performance means employees directly involved in transforming raw materials into goods or services which are specific to the job, the core technical skill. contextual performance concerns aspects of individual performance which not straight to their main function [5].

In this study using Factor Evaluation System method to analysing job performance, because there are many research about it [1,6,7]. Factor Evaluated System (FES has used two approaches i.e. point system and grade system. There are nine factors FES that common to most nonsupervisory position in general schedule occupations. To represent the intent of whole factor, several of FES factor have two or more sub-criteria. [6]. First to using this method the company must has been categorize all position become structural position or functional position. Functional position using classification standard FES. Structural position also called managerial position and to evaluated the job using classification standard FES plus classification to managerial.

The function of FES is to determine the value and class position that exist within the company. The value of position determines how much employee income, and class positions determine the duties and responsibilities held by a position. FES are public documents and should be available for review by anyone interested in their content. Furthermore there are nine factors in FES that not only evaluating duties and responsibilities but understanding about their position making valuation more detail. Thus making FES method widely used. The lack of the FES method is that method has not been able to deal with changes in factors that the company might consider important but not yet in the FES. A less detailed evaluated of employees activities, and FES does not have a standard value for each factor that must be obeyed by the position make the FES method less detail in analyzing job performance. Therefore the FES method is combined with the process mining. Until now there has been no research that discusses about FES method with business processes.

Process mining is a technique in the field of process management that support the analysis of business processes based on event logs [8]. Process mining aims at the extraction of information from a set of real executions (event log) [9]. Using company event log, it can be known that activities that really worked by employees because it has history of case, activity, time stamp and resources of employee [10]. Process mining helped FES method to analyzing the factor become more details. With the result that company can be easily to employee monitoring or give remuneration.

To make this method not only used for knowing the employee performance, so this method must have a weight for each factor. This paper used Analytical Hierarchy Process (AHP) to get the weight. AHP describes complex multi-factor or multi-criteria problems into a hierarchy [11]. With the weight, this method can be used too for recruitment and knowing the weaknesses of employee.

considering the advantages of FES and process mining, this paper proposes the combination FES and process mining to more detailing the factors of FES, which are required for recruitment, employee monitoring, training, remuneration and carrier planning.

II. RESEARCH METHOD

This proposed method is how to enrich FES method with combines the method and the process mining.

Step 1 – Collect all the company event log and standard operating procedure (SOP). SOP is a written set of instruction to be followed in order to achieve uniformity of performance of a specific function. An example of SOP is shown in Fig 1.

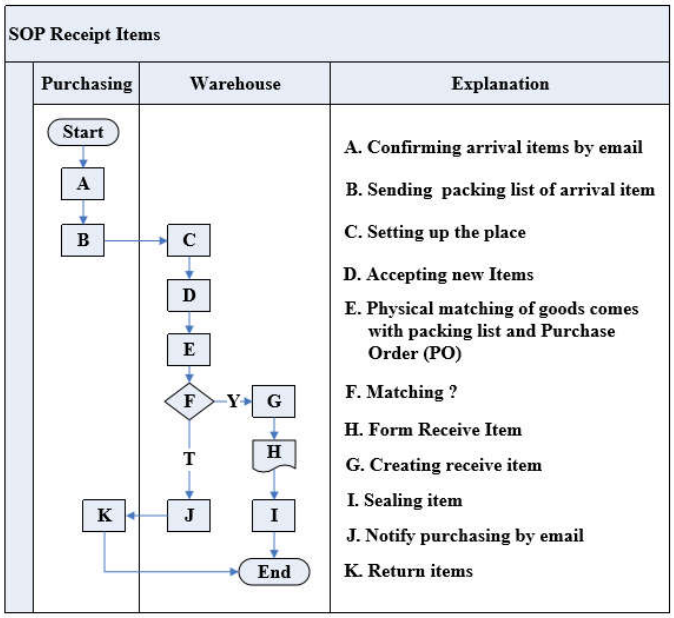


Fig 1. SOP Receipt Items

Table 1. Example of Event Log

Case ID	Activity	Start Timestamp	Complete Timestamp	Resources
01	Confirming arrival items by email	9/1/17 9:00 AM	9/1/17 9:06 AM	Purchasing
01	Sending packing list of arrival item	9/1/17 9:06 AM	9/1/17 9:09 AM	Purchasing
01	Setting up the place	9/1/17 10:00 AM	9/1/17 5:00 PM	Warehouse
01	Accepting new items	9/5/17 8:00 AM	9/5/17 2:00 PM	Warehouse
01	Physical matching of arrived items Purchase Order (PO)	9/6/17 8:00 AM	9/6/17 10:30 AM	Warehouse
01	Receive Item	9/6/17 10:30 AM	12/27/17 11:00 AM	Warehouse

Every position at least have one SOP and it linked with another position. So we must have the event log because it have some certain fields like case id, activity name, timestamp and resources. Table 1. is show the example attributes of the event in the logs. In that table there are case id, the activity name, start and finish time and the resources who did the activity.

Step 2 – From the SOP that have a Microsoft Visio format in Fig 1, drawing it again with YAWL format. YAWL (Yet Another Workflow Language) is an application used to create a business process model. It means, every process in SOP will be decomposed into detail activity based on the record logs. One the biggest challenge in this step is make a simulation models which represent the actual process.

Step 3 – Discovering the business process models. Purposes discovery is to find the process data and process model form which describe the actual activity. If the process model already fit then the next step is analysing the model process using classification standard FES. In the regular FES, each position will be evaluated based on 9 factors. Which is one position have more than one SOP. Table 2 is the description of FES factor.

Table 2 : Description of FES Factor

Factor	Name	Explanation
1	Knowledge Required by the Position	- Level of knowledge and skills are required - How to use skills and knowledge in the work
2	Supervisory Controls	- How to command task work - What is the employee responsibility if doing the work - How will the work be reviewed
3	Guidelines	- Type of guidelines used in doing the work - How much judgment are required to apply the guidelines or develop new guides
4	Complexity	- Character of the assignment - Level of difficulty in identifying what needs to be done - Inhibiting factor and authenticity in doing the work
5	Scope and Effect	- Work objective - Effect from work product
6	Personal Contact	- What is needed and the determination of where the contacts are made
7	Purpose of Contacts	- The reason “personal contact” is done
8	Physical Demands	- The use of physical force that is necessary in such work like frequency and intensity
9	Work Environment	- Physical environment in which the employee works and the specific safety rules or precautions in work

Every factor are not same but related to each other. Such as "purpose of contact factor" is sequel from the "personal contact factor". But every factor have different point rating. Table 3 is shown the point rating very factor.

Table 3. Rating and Point for FES Factor

Rating	Point FES Factor								
	1	2	3	4	5	6	7	8	9
1	50	25	25	25	25	10	20	5	5
2	200	125	125	75	75	25	50	20	20
3	350	275	275	150	150	60	120	50	50
4	550	450	450	225	225	110	220		
5	750	650	650	325	325				
6	950			450	450				
7	1250								
8	1550								
9	1850								

Rating in FES factor show the level of employee understanding from each factor. Rating higher means increasingly dominating on it job.

Step 4 – After analysing the business process model, then calculated every activities in position using standard FES.

Step 5 - Set the priority FES factor in every position using AHP. This step can be advantages for this research, because FES doesn't has standard value for every factors that must be obeyed by position. The basic step in the AHP method are :

1. Preparing a hierarchy of problems encountered
2. Make a pair-wise comparison matrix based on priority factors
3. Find lambda (λ)
4. Find Consistency Index (CI)

   Formula 1, CI aims to search the consistency of index. The good result is closer to zero.

$$CI = \frac{\lambda \text{ maks} - n}{n - 1}$$

Where :

CI = Consistency Index
N = number of criteria
= number of weight x factor

5. Find the Ratio Consistency Index (CR), CR < 10%

   Formula 2, If CR less than 0,1 so the result is consistent. If CR more than or as equal as 0,1 so the result is not consistent and the pair-wise comparison matrix must be repeated [12,13,14].

$$CR = \frac{CI}{IR}$$

Where :

CR = Consistency Ratio
CI = Consistency Index
IR = Index Ratio (Table 4.)

Table 4. Table of Index Ratio

n	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45	1.49

III. RESULTS AND ANALYSIS

Based on SOP in Fig 1. so the SOP re-drawn using YAWL.

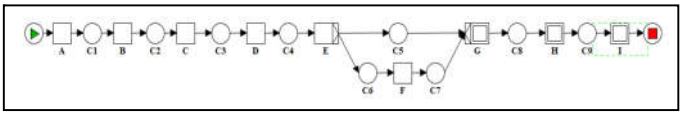


Fig 2. SOP Level 1 – Receipt Items

In Fig 2. There are 9 activity (from "A" until "I"). The powerful draw SOP with YAWL is the business process can be more details. There are 3 double box button that means that activity still have decomposition. Table 5 explain the notation in Fig 2.

Table 5. Explanation Notation SOP Level 1 – Receipt Item

Notation	Activity	Division
A	Confirming arrival items by email	Purchasing
B	Sending packing list of arrival item	Purchasing
C	Setting up the place	Warehouse
D	Accepting new items	Warehouse
E	Physical matching of arrived items Purchase Order (PO)	Warehouse
F	Notify purchasing by email	Warehouse
G	Received Item (2)	Warehouse
H	Sealing Item (2)	Warehouse
I	Entry Mc. Add (2)	Warehouse

From the description table 5. there are 2 activity which is not a warehouse officer jobs. And also have 3 activity that still have decomposition.

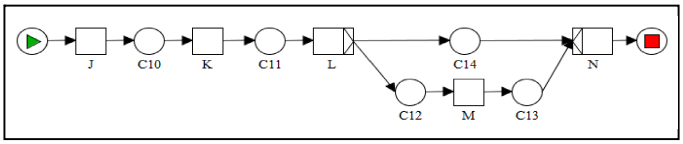


Fig 3. SOP Level 2 – Received Items (G)

Fig 3 describe the "G" notation in Fig 2. Actually "received item" consists of some more activities, but didn't have decomposition again. The description of Fig 3. is shown in the table below.

Table 6. Description Notation SOP Level 2 – Receipt Items (G)

Notation	Activity	Division
J	Creating New Form on Database	Warehouse
K	Select the Vendor	Warehouse
L	Approving selected PO	Warehouse
M	Edit the number of item received	Warehouse
N	Save Receive Item Form	Warehouse

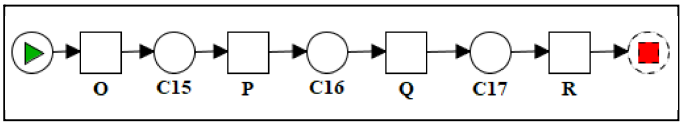


Fig 4. SOP Level 2 – Sealing Item (H)

Activity "Sealing Item" (H) in Fig 2. Explained in Fig 4. SOP level 2 – sealing item didn't have decomposition but the last activity in this process model will be continued in SOP level 2 – entry Mc. Add. The description of Fig 4. is shown in the table 7.

Table 7. Description Notation SOP Level 2 – Sealing Item (H)

Notation	Activity	Division
O	Writing the seal number	Sales
P	Open the item Packing	Warehouse
Q	Glue the seal	Warehouse
R	Entry Item Database	Warehouse

Activity "writing the seal umber" in warehouse SOP is turn out sales's.

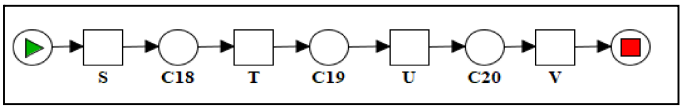


Figure 5. SOP Level 2 – Entry Mc. Add (I)

Fig 5. is the last decomposition on SOP Receipt Items. In this process model all activities are done by the warehouse. The Description of Fig 5 is shown in the table below.

Table 8. Description Notation SOP Level 2 – Entry Mc. Add (I)

Notation	Activity	Division
S	Open the Packing	Warehouse
T	Open the item database	Warehouse
U	Scan Barcode Mc.Add	Warehouse
V	Update item Database	Warehouse

After all business process model are made, so the next step is collect the warehouse activities. Every activity will be analyze using 9 FES factor. The following table is the collection of the activity from SOP Received Items Level 1 and Level 2.

Table 9. Activity in SOP Level 1 & Level 2

No	Activity	Resources
1	Setting up the place	indy, kholid, Hafid, Enggal
2	Accepting new items	indy, kholid, Hafid, Enggal
3	Physical matching of arrived items Purchase Order (PO)	Hafid, Bagas
4	Notify purchasing by email	Hafid, Bagas
5	Creating New Form on Database	Hafid, Bagas
6	select the Vendor	Hafid, Bagas
7	Approving selected PO	Hafid, Bagas
8	Edit the number of item received	Bunga, Hafid, Bagas
9	Save Receive Item Form	Hafid, Bagas, Migdad
10	Open the item Packing	Hafid, enggal
11	Glue the seal	Hafid, Bagas
12	Entry Item Database	Enggal
13	Open the Packing	Bagas, hafid
14	Open the item database	Bagas, hafid
15	Scan Barcode Mc.Add	Hafid, Bagas
16	Update Item database	Hafid, Bagas

Actually from all business process model in this research there are 19 activities. But 3 activity actually did by another position i.e purchasing and sales. So, we only just calculated 16 activity for warehouse officer position in SOP Received Items. If one position have more than one SOP just add up all the activity based on the point rating then divided by the total of activity.

FES method become powerful in analyze job performance. The basic FES just analyze a position by the 9 factors, but now the 9 FES factor already analyze 16 activities in a position. Table 10 shows how is FES method works. While have many factors, but FES can not know where are the employee weakness and stronger.

Table 10. FES method in a Position

No	Activity	Point
1	Knowledge Required by the Position	200
2	Supervisory Controls	125
3	Guidelines	125
4	Complexity	75
5	Scope and Effect	75
6	Personal Contact	60
7	Purpose of Contacts	20
8	Physical Demands	20
9	Work Environment	5

When every process in SOP is detailed become activities with process mining so the FES method become more detail. Table 11. Shows how are every activity in SOP Receipt Item from position warehouse officer calculated

Table 11. Activities with Factor FES in SOP Level 1 & Level 2

Activity / Factor	1	2	3	4	5	6	7	8	9	Total
C	350	125	125	150	75	25	20	20	5	895
D	50	25	25	25	25	25	20	20	5	220
E	200	125	25	25	25	10	20	20	5	455
F	200	25	25	25	25	25	20	5	5	355
J	50	25	25	25	25	10	20	5	5	190
K	50	25	25	25	25	10	20	5	5	190
L	50	25	25	25	25	10	20	5	5	190
M	50	25	25	25	25	10	20	5	5	190
N	50	25	25	25	25	10	20	5	5	190
P	50	125	25	25	25	10	20	20	5	305
Q	50	125	25	25	75	10	20	20	5	355
R	200	125	25	25	75	10	20	5	5	490
S	50	25	25	25	25	10	20	20	5	205
T	200	25	25	25	25	10	20	5	5	340
U	200	25	25	25	75	10	20	20	5	405
V	50	25	25	25	75	10	20	5	5	240
Total	1850	900	500	525	650	205	320	185	80	5215

Final score for this position obtained from total value for each activity then divided by total activity. Final score for warehouse officer is 5025 point divided by 16 activity equals 326. The proportional basic FES with the FES combined process mining is very different. Score for warehouse officer with basic FES is 705. It's almost doubled score FES with process mining, because its just count up the activities performed by employees.

The best score can be used to determine standard for new recruitment in that position. Using the event log in a period time, this FES method with process mining can analyze average and total time for one task. It's can be use by company to know the capabilities of employee who need training or reward.

For more detail in using FES with process mining, so added weight for each factor. In this research weight gets from the AHP method. Assign the weight start from asking the HRD or manager to determine the priority factor based on 9 FES factor. The factor start from the topmost until the small effect.

The next step is make a pair-wise comparison matrix. Number of columns are based on how many factors are used. Because FES has 9 factor, so we have matrix 9x9. It is shown in the table 12.

Table 12. Weight AHP for SOP Received Items

FES Factor	KR	SE	G	C	WE	PD	SP	POS	PC
Knowledge Required by the Position (KR)	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00	9.00
Scope and Effect (SE)	0.50	1.00	2.00	3.00	4.00	5.00	6.00	7.00	8.00
Guidelines (G)	0.33	0.50	1.00	2.00	3.00	4.00	5.00	6.00	7.00
Complexity (C)	0.25	0.33	0.50	1.00	2.00	3.00	4.00	5.00	6.00
Work Environment (WE)	0.20	0.25	0.33	0.50	1.00	2.00	3.00	4.00	5.00
Physical Demands (PD)	0.17	0.20	0.25	0.33	0.50	1.00	2.00	3.00	4.00
Supervisory Control (SP)	0.14	0.17	0.20	0.25	0.33	0.50	1.00	2.00	3.00
Purpose of Contacts (POS)	0.13	0.14	0.17	0.20	0.25	0.33	0.50	1.00	2.00
Personal Contacts (PC)	0.11	0.13	0.14	0.17	0.20	0.25	0.33	0.50	1.00

After calculate the pair-wise comparison matrix, the next step is count the weight from the average every factor.

Table 13. Weights of FES Factor

FES Factor	Weight	ƛ max
Knowledge Required by the Position (KR)	0.31	2.98
Scope and Effect (SE)	0.22	2.13
Guidelines (G)	0.15	1.50
Complexity (C)	0.11	1.04
Work Environment (WE)	0.08	0.71
Physical Demands (PD)	0.05	0.49
Supervisory Control (SP)	0.04	0.34
Purpose of Contacts (POS)	0.03	0.24
Personal Contacts (PC)	0.02	0.17

To test consistency of factor weight is gets from Random Consistency index (CR). The following are calculations using formula 1 and formula 2.

$$CI = \frac{9.60 - 9}{9 - 1} = 0.075$$

$$CR = \frac{0.075}{1.45} = 0.05 < 10\%$$

A Weight make a FES factor more powerful in analyzing job performance. This weight can shows the strength and weakness from the employee. The company can used the weight to standard for training employee to improve employee skill. But this weight can used to fired the lazy and can not improve employee. All the history recorded in event log and processed with FES method with process mining and AHP.

Table 14. FES with weight in activities

Activity	Total	FES Factor								
		1	2	3	4	5	6	7	8	9
		0.22	**0.15**	**0.11**	**0.3**	**0.04**	**0.02**	**0.03**	**0.08**	**0.05**
C	**895**	196.9	134.2	98.5	268.5	35.8	17.9	26.9	71.6	44.8
D	**220**	48.4	33	24.2	66	8.8	4.4	6.6	17.6	11
E	**455**	100.1	68.2	50.1	136.5	18.2	9.1	13.7	36.4	22.8
F	**355**	78.1	53.2	39.1	106.5	14.2	7.1	10.7	28.4	17.8
J	**190**	41.8	28.5	20.9	57	7.6	3.8	5.7	15.2	9.5
K	**190**	41.8	28.5	20.9	57	7.6	3.8	5.7	15.2	9.5
L	**190**	41.8	28.5	20.9	57	7.6	3.8	5.7	15.2	9.5
M	**190**	41.8	28.5	20.9	57	7.6	3.8	5.7	15.2	9.5
N	**190**	41.8	28.5	20.9	57	7.6	3.8	5.7	15.2	9.5
P	**305**	67.1	45.7	33.6	91.5	12.2	6.1	9.15	24.4	15.3
Q	**355**	78.1	53.2	39.1	106.5	14.2	7.1	10.7	28.4	17.3
R	**490**	107.8	73.5	53.9	147	19.6	9.8	14.7	39.2	24.5
S	**205**	45.1	30.7	22.6	61.5	8.2	4.1	6.15	16.4	10.3
T	**340**	74.8	51	37.4	102	13.6	6.8	10.2	27.2	17
U	**405**	89.1	60.7	44.6	121.5	16.2	8.1	12.2	32.4	20.3
V	**240**	52.8	36	26.4	72	9.6	4.8	7.2	19.2	12

Table 14 is the weight for each activity obtained from the value activity in table 11 that multiplied by the weight of FES factor. In table 11 the factors that have passed the standard are marked in red.

Considering the advantages of FES and process mining, this paper proposes the combination FES and process mining i.e :

1. Employee Monitoring

   The company can use details of activities to notice deviation that maybe caused by employee. Based on FES which includes factor to monitor the employee are supervisory control (2), guidelines (3) and personal contacts (6).

2. Remuneration

   If the result of monitoring is good, so the employee have a right to get reward or remuneration. Complexity (4), scope and effect (5), and physical demands (8) are FES factor that affect to give it.

3. Employee Training

   Training of employee is a must to increase a job performance in company. With this method, the company can find which part of the employee who need training. Which became the basis for decision-making training on FES is knowledge required by the position (1), purpose of contacts (7), personal contacts (6) and "work environment (9).

4. Recruitment

   With the help of AHP , the combination of FES and process mining became more powerful. FES not only use for analyze job performance but for recruitment new employee because all the factor in FES have standard.

IV. CONCLUSION

Based on research, this paper proposed a new method for job performance analyzing using Factor Evaluation System and process mining. Before this paper, usually company just using FES to analyze job performance. And this paper using AHP to get weight for each factor. This weight aims so that company not only just using this method for remuneration and career planning, but for recruitment and planning too.

The work was tested on a real-life event log data which are taken from an organization which has a focus on selling and distribution fiber optic. The result shows that this offered method is more details in analyzing job performance i.e employee monitoring using factors 2,3,6; remuneration using factor 4,5,8; employee training using factors 1,6,7,9; recruitment using all factor which have weight standard.

For future works, the standard score for every factor must be arrange according the company needs.

References

[1] L.J. Tiesinga, MSN RN R, J. G. Halfens, phd , J.T. Algera-Osinga MSN RN and A. Hasman, phd. The Application of a Factor Evaluation System for Community Nursing in the Netherlands. *Journal of Nursing Management*, 1994, 2, 175-179.

[2] XInbo Sun, Neng Luo. Study on the Effectiveness od Point-Factor Job Evaluation System in Operation Position. *Journal Communications in Information Science and MAnagement Engginering,* 2013, Vol. 3 Iss. 3, PP. 154-160.

[3] A. Iskandar, E. Salbiah, I. Purnamasari. Analysis of Structural Position Performance With Factor Evaluation System (FES) Reference To Determine Work Achievement Indicators With Remuneration As Intervening Variables in Bogor, West Java Province, Indonesia. *International journal of Sciences : Basic and Applied Research (IJSBAR),* 2014, Vol. 18, No 2, pp 223-234.

[4] Korkaew Jankingthong* and Suthinee Rurkkhum. Factors Affecting Job Performance : A Review of Literature. *Silpakorn University Journal of Social Sciences, Humanities, and Arts.* 2012; Vol.12 (2) : 115-127.

[5] Janetta Sirbu, Florin Radu Pintea. Analysis and Evaluation of Jobs – Important Elements in Work Organization. *Procedia – Social and Behavioral Sciences,* 2014, 59-68.

[6] Office of Personnel Management (OPM) Classification Standards. TS-107 August 1991. The Classifier's Handbook. U.S Office of Personnel Management; 1991.

[7] Peraturan Menteri Negera Pendayagunaan Aparatur Negara dan Reformasi Birokrasi Republik Indonesia Nomor 34 Tahun 2011 Tentang Pedoman Evaluasi Jabatan. PermenPAN dan RB 34 Tahun 2011. Negera Pendayagunaan Aparatur Negara dan Reformasi Birokrasi; 2011.

[8] Riyanarto Sarno, Fernandes Sinaga. Business Process Anomaly Detection using Ontology-based Process Modeling and Multi-level Class Association Rule Learning. *International Conference on Computer, Control, Informatics and its Applications (IC3INA)*. Pages: 12 – 17, 2015. DOI: 10.1109/IC3INA.2015.7377738 .

[9] Riyanarto Sarno; Hari Ginardi; Endang Wahyu Pamungkas; Dwi Sunaryono. Clustering of ERP business process fragments. *International Conference on Computer, Control, Informatics and its Applications (IC3INA)*. Pages: 319 – 324, 2013. DOI: 10.1109/IC3INA.2013.6819194.

[10] Wil M.O. van der Aalst. Process Mining : Discovery, Conformance and Enhancement of Business Processes. ISBN: 978-3-642-19344-6, DOI 10.1007/978-3-642-19345-3, Springer Heidelberg Dordrecht London New York, 2002

[11] Nachiappan Subramanian and Ramakrishnan Ramanathan. A review of applications of Analytic Hierarchy Process in Operations Management. *International Journal of Production economics*, 2012; Vol. 138 (2) : 215-241.

[12] Afrianda Cahyapratama, Riyanarto Sarno. Application of Analytic Hierarchy Process (AHP) and Simple Addictive Weighting (SAW) Method in Singer Selection Process. *International Conference on Information and Communication Technology (ICOIACT),* 2018. DOI: 10.1109/ICOIACT.2018.8350707.

[13] Uky Yudatama, Riyanarto Sarno. Priority Determination for Higher Education Strategic Planning Using Balanced Scorecard, FAHP and TOPSIS (Case study: XYZ University). *IOP Conference Series: Materials Science and Engineering* (Vol: 105, Issue: 1, Page: 012040). DOI: https://doi.org/10.1088/1757-899x/105/1/012040.

[14] Uky Yudatama, Riyanarto Sarno. Evaluation Maturity Index and Risk Management for IT Governance using Fuzzy AHP and Fuzzy TOPSIS (Case Study Bank XYZ). *International Seminar on Intelligent Technology and Its Applications (ISITIA),* 2015. https://doi.org/10.1109/isitia.2015.7220000

Media Economics and the Use of Technology in the Production of Local Gaming Industry

(A Study at Creacle and Wisageni Game Studio Yogyakarta)

Muhamad Hasan Basori[1], Lisa Mardiana[2], Wildan Namora Ichsan Setiawan[3]
Department of Communication Science
Dian Nuswantoro University
Semarang, Indonesia
[1]hasansbasori@gmail.com, [2]lisa.mardiana@dsn.dinus.ac.id, [3]wildannamora@gmail.com

***Abstract*— This research aims to reveal how the dynamics of the local game industry, especially in Yogyakarta. To be able to survive in the market and compete with foreign game products, the local game developers must also be able to adapt to the latest technology, to produce the games that the consumers want and have unique characteristics, because the game is one of the media industries that also has and uses limited resources to produce entertainment contents in digital forms that is distributed to the consumers to satisfy their needs. This research tries to examine the structure of the local game market players, and how the production process carried out with the technology is used optimally. The results show that in Yogyakarta, this type of industry has increased along with the economic growth, and on the stage of production process, the game developers use the technology with the game engine whose output can be applied to a multiplatform called Unity.**

***Keywords*—gaming industry, local game, video game, game engine**

I. INTRODUCTION

Before the technology develops rapidly as now, in Indonesia, the games, especially the traditional ones, can only be done by individuals or just a few people, and they also must play face to face and have enough space to play together, for example sack racing, marbles, jumping rope, card games and so on. In this advancement of communication technology era, various sophisticated devices have been created, such as PCs, Tablets, Smartphones and so on. So, the game, in its development process, has evolved to become more massive in its use, using communication technology connected in the internet network (online). The game can be played alone or only by two or more people at once in real time and does not require open any space to play because it is now in a form of digital content that is played in cyberspace. Esposito [1] emphasized that the technology has been widely used in playing video gaming activities, the technology in the form of electronic devices with the capabilities of computer and digital technology, and software in the form of applications or programs which then form the gameplay, and the story line in a video game.

The game has changed as one of the big industries, becoming a commodity of the needs of contemporary society which is influenced by the culture of capitalism. The profits obtained by the mobile game publishers in 2013 reached 17.5 billion USD, in 2014 its increased by 25.5 billion USD and in 2015, 30.3 billion USD. Then, it is expected to increase again to USD 40.8 billion in 2017 [2]. The data are accumulated from the profits of mobile game sales on smartphones and tablets. Distribution and storage of entertainment products through digital format as the development of new media, especially smartphones, results in direct access to online distribution channels for the game publishers, making it easier to market their mobile game products.

In a market-based economy, supply and demand interact to make the economic process works. Mass media continuously puts itself in the mechanism of supply and demand, managing daily-based resources to supply the consumers with media/product contents. Albarran explained three main concepts in the media economics, i.e. resources, production and consumption [3]. So, to be able to compete in the game market, the local gaming industry must also be able to adapt to the market trends, especially in managing various resources, production processes, and targeted consumers.

Several early local Indonesian game developers in the gaming market, Altermyth, Algerium, Artlogic Games, Own Games, Menara Games and Toge Production have produced unique characteristic games that can compete with imported ones. Toge Production creates games in some genres with funny visual character graphics using pixel art techniques for the mobile games. Toge is famous from two game titles on the desktop gaming platform, namely: Infectonator: World Dominator, and Planetary Conflict, which have attracted more than 159 million users.

Likewise, Creacle and Wisageni, as the local game developers in the Yogyakarta, also have transitioned their game products to mobile platforms. But, Creacle continues to manage the products on desktop platforms, depending on client requests, unlike Wisageni, which only focuses on managing platform mobile products. The author studies the production process carried out by Creacle and Wisageni,

because the production process becomes a crucial factor in the success of products on the market, thus successfully increases opportunities and is able to make good profit.

Creacle and Wisageni game studios are also chosen as the research objects for three reasons. First, from the historical background, Creacle is arguably one of the pioneers of the gaming industry in Yogyakarta, so it surely knows more about the conditions of the gaming industry than Wisageni which is still a start-up company. Secondly, the game design is unique. Although Wisageni is still classified as a new player in the gaming industry, but the game titles that have been created and released are warmly received by the consumers due to different themes it offers, while Creacle which is has been well known by the local game industry can also survive with unique game designs. Third, they have the same game type and platform, Creacle and Wisageni are two developers which create casual games. The Creacle and Wisageni products can be accessed and played on a mobile or smartphone platform.

This research only focuses on the production stages in the local gaming industry and the use of technology with the Unity game engine to create a game. Based on the background described, the problem of this study is formulated in the following question: how is the production process and technology of game applications in the local game developers, Creacle and Wisageni Game Studios?

II. Literature Review

A. Media Economics

Albarran [4] defined the media economics as a study of how the media industry uses limited resources to produce services distributed to the consumers in the community to fulfill various desires and needs. Based on this definition, it can be seen that there are three main concepts in the media economics, i.e. resources, production and consumption. The resources, in economic term, are things that are used to produce goods and services. This includes intangible things such as human resources, cameras, etc. The production is a series of activities to create various items for consumption (for example, the production process of printed media, electronic media, film, recording, books etc.,), and the consumption is the use of goods and resources to fulfill desires and needs (the use of various types of media or markets).

B. Steps of Game Production Process

According to Les Pardew, the game development is divided into three stages: pre-production, production and post-production [5]. Each stage is further divided into sections. In detail, the game development stage can be explained as follows: a) pre-production is an important stage because it becomes the foundation of the entire game making process. In this early stage, the game designer formulates the game concept through depictions from the concept of ideas to the prototype; b) production is the stage where the game that is going to be developed into a product that is ready to be marketed. In this stage, a lot of works are done, including making all game assets, implementing all assets into the game, and conducting second trial on the prototypes, and (c) post-production is the final stage in the game development process, where there will be beta test, then duplicated and packaged as attractive as possible.

C. Game Engine

A game engine is software that provides a set of functions that are required in creating games quickly and efficiently for the game makers [6]. Since 1993, it initially only supported for 2D models. After 25 years of development, this engine has fully supported 3D models, lifelike images, massively multiplayer online games, AI (artificial intelligence) and mobile platforms [7]. The game engine framework now supports the game development by integrating some areas such as art import and assets, 2D and 3D, collecting them into scenes and environments, adding lighting, audio, special effects, physics and animation, interactivity as the part of digital era, game logic, editing, debugging, and content optimization for the target platform [6]. One of the most popular game engines in this industry is Unity. Unity provides various resources that can support the creators in building and developing the games with deployment support on more than 25 platforms.

III. Research Methods

This study used a qualitative approach, and the selection of informant used purposive sampling. The informant in this study was Gathot Fajar as the CEO and the Game Designer of Creacle Game Studio, and the interviewed informant at Wisageni Game Studio was Rudy Sumarso as the Creative Director and the Wisageni Game Designer. The primary data were obtained from in-depth interviews, and the secondary ones were obtained from document archives, and online sites in the form of data from digital market platforms, and the results of the research on video game markets in various countries, and online newspapers. The data analysis technique used was qualitative taken from Miles and Huberman which consisted of three activities that were done simultaneously: data reduction, data presentation, conclusion drawing / verification [8].

IV. Result of The Research

Yogyakarta is a region with abundant creative economic potential, with many quality universities, as well as high interest from industry players, making it a city with a very healthy game development ecosystem. Games are also one of the leading sectors, besides movies and music which can be a compelling factor for other sectors. The gaming industry in Yogayakarta has been developing well since several years ago. Most of the developers in Yogyakarta are indie, coming from a group of teams that work purely on the basis of passion for video games, and strong intentions to advance the domestic game industry. Indonesia accounts for half of the video game market amounting to 1.4 billion USD in the Southeast Asian market, and the Indonesian gaming market alone can generate revenues of 321 million USD in 2015.

The revenue was a 56.0 increase compared to 2014, making Indonesia placed 24th rank in the world's largest video game market, and the second in Southeast Asia. From 42.8

million gamers in Indonesia, 24.1 million spend their money on mobile games, and the ratio of player payments is more than 56%, above the regional average [9]. Therefore, the gaming industry in Indonesia has developed rapidly in recent years and it will be brighter in the future. Besides, recognition and awareness of the public is increasingly open about the domestic video game industry, making good opportunities for the local game developer studios in Yogyakarta.

A. Structure of Local Gaming Industry in Yogyakarta

Varian (1993) described that microeconomics focuses on the study of the activities of components of economic systems such as markets in a specific scope, companies and consumers [10]. The microeconomics analyzes the structure of the market, how it works, and how it behaves. On the national scale, the market structure of the gaming sub-sector is highly developed, with the existence of more than 80 indie game developers. The dynamics that arise between these actors have more or less changed the game market landscape into perfect competition but unclear structure. There are six local players in Yogyakarta, and the industrial growth trend increases along with economic growth. Although there have been many game titles created by local game players in Yogyakarta, but from product differentiation factors based on character and type of game products, they tend to be the same as the casual games. The product differentiation in this industry is still relatively low due to less diverse products.

The mobile games have become a trending commodity platform and are popular in the global gaming industry, especially Indonesia. The local game developers lead to mobile phone platforms because of the process of creating alternative game contents quickly and economically. It is in line with Juul who stated the fact that cell phones can propagate and attract non-traditional gamers in the large market of casual mobile gaming platforms [11]. Besides, the casual gaming market targets male and female users with diverse demographic backgrounds of all ages.

1) Creacle Game Studio

Creacle was established and used as an identity in August 2011. Over the time, in 2012, Creacle began to be registered itself as a creative CV. Creacle, which stands for Creative Circle, means that the circle represents something unlimited, has no angle or boundary, so that Creacle hopes to become a company that has unlimited creativity like the logo form that comes from the basic shape of the circle. Creacle Game Studio has spawned more than 30 video game titles officially released on various platforms. The game titles include Connect Smabung Kata, Kompilasi 17an, Ghost Academy, Tebak Gambar Pahlawan, and others. The structure of the production at Creacle is divided into three divisions, namely the CFO (Chief Finance Officer) managed by Eko Nugraha in charge of administrative affairs, CTO (Chief Technology Officer) managed by Irkham Huda who is in charge of game development, and COO (Chief Operational Officer) managed by Rifauddin Tsalitsy in the operational section. The three sections are under the supervision of Gathot Fajar as the CEO. Creacle Game Studio has received a number of awards on the game titles, including: "Save Komo" received an award at the 2012 Indonesia Game Show, then "Virtual Pet" as the first winner in the Nokia Game Dev War 2, and "Green Guard" which gets an award for the mobile game category at INAICTA 2011.



Fig. 1. Logo Creacle Game Studio [12]

2) Wisageni Game Studio

Wisageni Game Studio is one of Yogyakarta's local game developers which was established around November 2014. Initially, Wisageni created web games for browser users, but there was a change in market interest around August 2016 to the mobile games market, because the web game market experienced a significant decline in revenue, not only in the local gaming market but also at the global level. Wisageni is a game studio in Yogyakarta that has a dream to give a new color to the game industry in the world with the uniqueness of Indonesian culture. As the name implies, Wisageni itself is taken from the philosophy of the name of the puppet character, Wisanggeni. Wisanggeni's character is purely a Javanese and it is not included in Mahabarata epics from India. Wisageni's organizational structure is divided into three: the Creative Director and founder is Rudy Sumarso, as a Producer and co-founder is Vania Marita, then as Technical Director and also co-founder is Dhimas Prabowo. The awards that have been achieved by Wisageni include "Best Gameplay" nomination in the 2015 InGAME organized by Gamelan, winning 1st place in the category of game in the Increfest (Indonesia Creative Festival) 2015 organized by ministry of Communication and Technology, the Finalist in the AYCIF (ASEAN Youth Creative Industry Fair) 2015, and IndiePrize Finalist Award in Casual Connect Asia 2015.



Fig. 2. Wisageni Game Studio [13]

B. Process and Stages of The Game Production

In the media economic activities, there is a scope of vertical integration which contains the integration, synergy, or cooperation in various aspects of the organization, such as production, distribution and exhibition. In shorter words, vertical integration is a media company that creates products from upstream to downstream, because the communication technology cannot be separated from the industrialization process. All economic reasons in the industry get the tools to further strengthen their economic position in technology [4]. The process in the game industry is called the game development, and each stage of the game development process in Creacle and Wisageni will be described as follows.

1) Pre-Production

This pre-production stage is the main task for game designers. On this stage, the game designer starts to do analysis which includes market research, trends, and what kinds of games can be accepted by the market. This stage is to process the basics of all elements that will be included in the game, starting from Mechanics, Stories, Aesthetics and Technology. At Creacle, the main players of this early stage are Game Designers and Programmers. The Game designers conduct market research, research trends, and what games the market can accept, and write gameplay design plans, mechanics, and levels of the game. The programmers can design a User Interface model to get a 'base skin' design as the first stage design. The output produced in this pre-production stage is the Game Design Document. Creacle has two product divisions, and there are also two types of Game Design Documents, namely the High Concept document for the product division ,and the Pitch document for the service division.

Although Wisageni has ideas or themes that are different from other game developers, but the pattern of the game development process is quite similar. This pre-production process can be called the design phase because it is done by the creative director who is also a game designer, and this phase is also the project planning that emphasizes the development of ideas or concepts and the creation of initial design documents. The purpose of concept development is to produce documents that describe all activities, schedules and development teams clearly and easily understood for the entire game development team. The difference is Wisageni does not serve clients. Wisageni design documents are called "Concept documents". These documents are the concepts that contain a more detailed explanation of the game that will be developed when compared to the Pitch document. Other components include game flow and descriptions of how to play, features, stories, target markets, hardware platforms, estimated schedules, marketing analysis, team requirements, and risk analysis.

2) Production

The production stage in the game development process is a series of steps that require capital, resources and long time, depending on the complexity of the game created, and sub-stages are described as follows:

At Creacle, the main actors on this stage are Game Designer, Programmer, and Game Artist. First thing to do is to arrange project management by the managers coordinated by each division, so that the game development process is in line with the target time, budget allocation, and desired quality. Secondly, the development and design adjustments stage; this stage is an important and collaborative process to create contents and rules in the designed game. The game designers act as the parties who monitor the development and adjustment of the game designs to match the high concept of the games to be created. The third is the stage of making game assets. At Creacle, the game artists begin to make concept art to be used as a temporary asset for the programmers to start designing the gameplay. On this stage, the programmers, using the game engine called Unity, have completed assets that were previously lacking, also have completed the games with sound effects even though not yet perfect. All of these can be seen as a whole game created. The fourth is the stage of game testing. It is important to know if there is less smooth mechanisms process in the game.

At Wisageni, the main actors of this stage are Creative Director, Producer and Technical Director. Although the designation is different, the role is the same as in Creacle. First, the project management stage in Wisageni is managed the Creative director. This stage is to execute plans in the design document. The parties who play the roles in this stage are also the producers and the Technical Director that must coordinate well each other. Secondly, on the stage of developing and adjusting the document design, it is the duty of the creative director as the game designer, which is responsible for developing and adjusting the game design to fit the concept document of the game to be created. Third, the game asset creation also works on audio creation done by the producers. The audio in the games is grouped into three categories: sound effects, music, and voice-over.

The audio here doesn't have to be done alone, because the audio and music can also come from outsource. Fourth, the trial phase is also significant in the success of a game product because this stage is needed to guarantee the quality of a product. The trial conducted by the game tester is brought in through the announcement delivered by Wisageni via email or social media.

The output generated from this first stage of production is the first playable. The goal is to find out the weaknesses and start the components being repaired, and making newer version until it goes into the alpha stage. On this stage the asset has been relatively complete and can be seen as a whole of the features of the game created. In the end, Beta testing aims to get input and criticism as a benchmark for polishing the game that has been tested for the third time. After the beta phase is over, a newer version is made, until it can be called as “a game” that can be used as a release version. Creacle itself applies the minimum version or called Minimum Valuable Product (MVP), which is the version where the game has been assessed to be eligible for release.

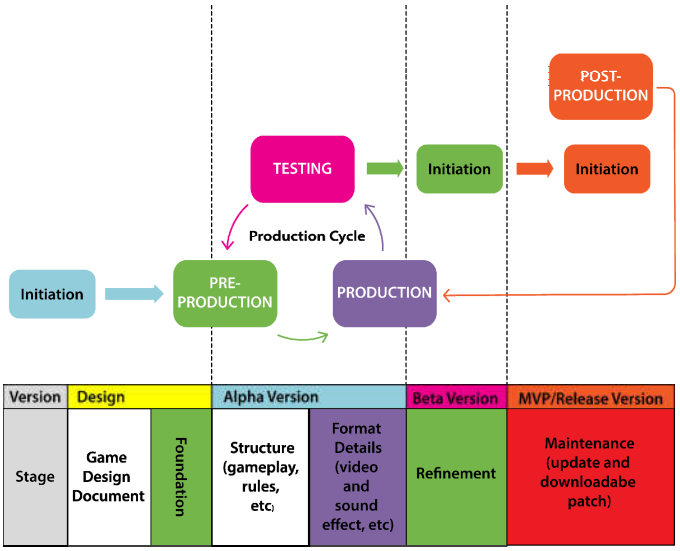


Fig. 3. Cycle and Process of Game Production

3) Post-Production

This process is not the end of game development, because this final process is also the stage of improvement or renewal. Because there is no perfect game, the game titles released by Creacle are always waiting for input through reviews, so that updates and maintenance can be done periodically. As with other game developers, the post production at Wisageni which is included in the game development on mobile game platforms can maintain the released games by providing downloadable patches.

The output from the post-production stage is the final product in the form of game master released in the market. The final distribution of the game products to the market or to consumers can be done through several retail / platforms such as social media, mobile gadgets such as the Android Market, and Google PlayStore. The distribution can be done by game developers and publishers. This is because there are some games that can be distributed without any help of the publishers.

C. Use of Technology (Unity Game Engine)

Basically, technology is an object whose development is always dynamic and difficult to predict its progress, and used to facilitate all human activities. Since the humans are getting to know technology, there have been many technological inventions created to this day, and some of them are used to create new types of games. The earliest games were created in 1947, and they were initially deployed for the simulations missile launch on specific target. The games that were originally created only for military purposes, have turned into one of the popular entertainment media, lifestyles and human livelihood objects.

Technological advances are very influential for the game developers. "Technology" is a medium to realize the games they want to create; technology is what makes a vague or impossible idea to be possible. The game developers have a job that is arguably heavy, and the job is to be able to create new games periodically, because people buy or play new games simply because they are new. Pressures such as new things and the sophistication of new game play devices (the emergence of new computer hardware, new game consoles, new smartphones and others), are factors that make the game developers can no longer only think about their goals to create extraordinary games, but also about what game engine technology should be used to create games that can be played on new game platforms used by the players.

The game engine technology called Unity becomes the main choice for many local game developers, especially those who create mobile games such as Creacle and Wisageni game studios. Using Unity, designing game objects and giving functions or commands to cold objects is easier. In each project of Unity, there are several variables that must be understood, including: asset is a storage place for all components of the model, texture, image and audio files. Scene is a visual display that contains the game contents, such as game interface, game level, game menu and so on. Game objects are components in assets that can be moved into the scene to be moved, rotated, and arranged in certain composition. Script features provided by Unity support 3 programming languages, namely JavaScript, C #, and BOO.

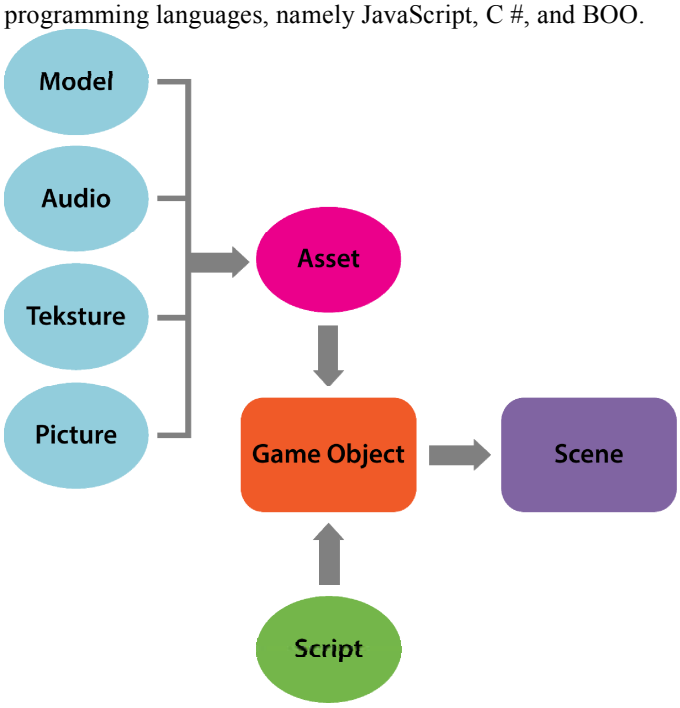


Fig. 4. The Unity Component Model

V. Conclution And Suggestions

In accordance with the media economic paradigm used in this study, it is found that the production process in the gaming industry has a complex set of stages in the game development cycle. The long production process does not only create entertaining contents, but prioritizes economic motives, and the successful game products are those created with good monetization strategies. Thus, the results of this study prove that Creacle's production process is casual and trivia, while

Wisageni produces games that are also casual and collectible. Both game developers create the games with first person game mode by importing 2D and 3D images which are then used as the assets for the object games in Unity game engine.

Suggestions for Development

The economic media perspective can be used on the products of the gaming industry by looking at the characteristics of the game products that have been successfully released to the market, because the contents of game products are not just created, but have gone through a long process from a series of production steps that are not easy and take times. From this study, the author finds a fact that the local Indonesian game industry is able to compete in the market viewed from its products, but still has some weaknesses in the economic monetization strategies applied in the systems and use of technology, as expressed by Gathot as the CEO of Creacle who has been in the business for a long time:

"*There are many game designers having good story ideas, good description of concepts, but many of them get stuck in the monetization section. Because seen from their experience, the people now have been able to make games, but how to make money from it, not so many think in detail about it. There have been many games created, but to make the good ones and make money, it is still a serious problem for us.*"(Interview with Gathot, February 15, 2017).

So, it requires direction from stakeholders and also the government in determining the policies of supporting ecosystem development and the use of game engine through clear roadmap for the creation of a local market with global quality that benefits the local industry.

Acknowledgment

The authors would like to thank to thank Dian Nuswantoro University for supporting this research program.

References

[1] N. Esposito, "A Short and Simple Definition of What a Videogame Is", University of Technology of Compiegne, 2005.

[2] T. Wijman. (2017, November 28), "Global Games Market Will Reach 102.9 Billion In 2017". [Online]. Available: https://newzoo.com/insights/articles/global-games-market-will-reach-102-9-billion-2017-2/

[3] Usman Ks, "Ekonomi Media: Pengantar Konsep dan Aplikasi". Ghalia Indonesia, 2009.

[4] A. B. Albarran, "Media Economic: Understanding Markets industries and conscepts", USA, Iowa State University Press, 1996.

[5] L. Pardew, "Beginning Illustration and Storyboarding for Games", USA, Thomson Course Technology PTR, 2004.

[6] Unity Technology, "Game engine-how do they work?", 2018. [Online]. https://unity3d.com/what-is-a-game-engine

[7] J. Xie, Research on key technologies base Unity3D game engine. In: ICCSE 2012 - Proceedings of 2012 7th International Conference on Computer Science and Education, pp. 695–699 (2012)

[8] M. B. Miles, A. M. Huberman, Analisis data Kualitatif, Penerjemah: Tjetjep Rohendi Rohidi. Jakarta, Universitas Indonesia, 1992.

[9] "Southeast Asia Games Market_The World's Fastest Growing Region_Casual Games Sector Report 2015", Newzoo, [Online] Available: https://s3.amazonaws.com/CGA_Report/southeastasia-report-2015

[10] H. R. Varian, "Intermediate Microeconomics: A Modern Approach. Third Edition", New York: W.W. Norton & Company, 1993.

[11] J. Juul, "A Clash between Game and Narrative: A Thesis on Computer Games and Interactive Fiction" (2001 English translation from Danish), Institute of Nordic Languages and Literature, University of Copenhagen, 1999/2001.

[12] Creacle Studio (2013, March 6), "Foto Profil". [Online]. Available: https://www.facebook.com/creaclegamestudio/photos/a.266640366692069/531303990225704/?type=3&theater

[13] Wisageni Studio, "guy.jpg", 2017. [Online]. Available: http://www.wisageni.com/assets/images/guy.jpg

Twitter Sentiment Analysis of Movie Reviews Using Information Gain and Naïve Bayes Classifier

Sari Widya Sihwi
Informatics Department, FMIPA
Universitas Sebelas Maret
Surakarta, Indonesia
sariwidya@staff.uns.ac.id

Insan Prasetya Jati
Informatics Department, FMIPA
Universitas Sebelas Maret
Surakarta, Indonesia
prasetyainsan@student.uns.ac.id

Rini Anggrainingsih
Informatics Department, FMIPA
Universitas Sebelas Maret
Surakarta, Indonesia
rini.anggrainingsih@staff.uns.ac.id

Abstract— **The opinion of Twitter users related with movie is a valuable asset to be studied. This is because business and organization casts always want to know the public and consumers' opinions about their movies and consumers also want to know others' opinion before purchasing movie products. Sentiment analysis was applied in this research to gain information whether a tweet is positive opinion, negative opinion, or neutral opinion. Naive Bayes Classifier was chosen as the algorithm due to its strength on accuracy. In order to increase the run time efficiency, it was combined with Information Gain as feature selection method. By collecting tweets from 12 popular movie titles as dataset, the experimental evaluations show that the proposed techniques have the accuracy 82.19% with 0.006 as the optimal threshold of gain.**

Keywords— information gain, naïve bayes classifier, sentiment analysis, twitter

I. INTRODUCTION

The development of internet has changed the users' behavior in expressing thought, feeling, and interaction with social environment. The behavior pattern produces data in big amount and brings us to the birth of data mining as the subject discipline playing a role to extract the information which is not known previously. Data mining can be said as a natural evolution of the information technology and a companionship of some subject disciplines and related application fields [1]. Twitter is social media mostly interested in the data mining process compared with other social media because of its easy-to-use and popularity factors, that it is able to produce data in big amount with large topic variations [2]. The easy of use which is very beneficial is an open data access with the feature of Twitter API [3].

The opinion of Twitter users is a valuable asset to be studied. Business and organization casts always want to know the public and consumers' opinions about their products and services. Consumers also want to know the users' opinions of certain products before purchasing, or others' opinions about politic candidates before deciding to choose in the general election [4]. Therefore, the Twitter data have been used for predicting the various phenomena such as crime, stock exchange, general election, opinion probe, public health, and movie purchasing [5]. The activity of opinion data processing is one of the data mining which is called as opinion mining or sentiment analysis.

Sentiment analysis is aimed for getting to know whether the opinion is positive, negative, or neutral [6]. It can be said as the combination between text mining and natural language processing [7]. It is a research branch studying about opinion, sentiment, evaluation, appraisal, response, and emotion of the people towards various entities such as product, service, and other related matters [4], which is needed in the process of decision-making because "What other people think" always becomes an important part [6]. This is the reason why getting public opinion has, a long time, been a big business for marketing needs, public relations, and politic campaign institutions [4].

One interesting content or product for sentiment analysis is movie. Based on the research issued by Motion Picture Association of America (MPAA), the data of the total movies produced in 2015 are 791 movies from movie studio which either becomes member or non-member of MPAA [8]. A big number of movies make the consumers become more selective to decide the movie being watched. The movie review has an effect on the consumers' thinking process which consequently leads to a decision whether or not it is necessary to buy a ticket to watch a movie. Generally, actually consumers trust more on the other consumer appraisals than the critics' in giving the accurate information. This is because they think that movie consumer fellows have more similar taste and understanding [9].

The data collection and opinion research from the web become a need, moreover with the growth explosion of the social media such as online discussion forum, micro blog, Twitter, comments, and posts in the social media [6]. The application of sentiment analysis has been widely spread in all fields, from the consumption of products, services, health, finance, social activities, and general election in the politic area [4].

The tendency shows that the opportunity for consumer's review exploration is very open that there are so many unstructured data in the social media which potentially enrich the insight of the consumer's preferences towards a movie. The sentiment analysis of Twitter consumers is an alternative solution to find out the information.

The modern approach which is used in the sentiment analysis is machine learning technique which is able to build classification model by training some labeled document sets. Sabeeha Mohammed Basheer did an experiment to identify the best machine learning method in the the sentiment analysis approach. The result, Naïve Bayes Classifier is the most appropriate approach with high accuracy and quick run-time [10].

One problem in the sentiment classification is if there are too many attributes used in the dataset then the accuracy getting lower [11]. The Naïve Bayes Classifier method itself also have the same weakness which is sensitivity to the too many attributes or features that it reduces the accuracy level [12].

The weakness can be overcome by using feature selection algorithm which can increase the accuracy level and the time efficiency of the classifier. Dandhe and Patnaik, in their research, studied about sentiment analysis from movie review dataset using Naive Bayes Classifier and one of the feature selection algorithm, Unigram [13]. The performance of the two methods only reached the accuracy of 62.35% and showed that Unigram had not been able to increase the classifier accuracy optimally. Chandani and Wahono compared five most appropriate feature selection algorithms for the sentiment analysis on the movie review dataset [11]. The five algorithms were information gain, chi-square, forward selection, and backward elimination.

The information gain algorithm gained the highest average classification accuracy with percentage of 84.57%. Thus, the Information Gain can be concluded as the best feature selection algorithm in the sentiment analysis. Therefore, Information Gain becomes the chosen method to be feature selection algorithm.

II. NAÏVE BAYES CLASSIFIER

Bayes classification method is well-known as Naïve Bayes. This method assumes that the effect of attribute value in certain class is independent to the other attributes. The assumption is named class-conditional independent which makes this classification method believed as naïve [1]. The basic theorems of the Bayes method is written as follows [1] :

$$P(H|X) = \frac{P(H)P(X|H)}{P(X)} \quad (1)$$

The application of the Bayes theorems as the machine learning method in the classification method is known as Naive Bayes Classifier method.

III. METHODOLOGY

The methodology being planned in this research is explained in the fig. 1. All the stages in this research are using PHP programs language: Hypertext Processor (PHP) and MySQL.

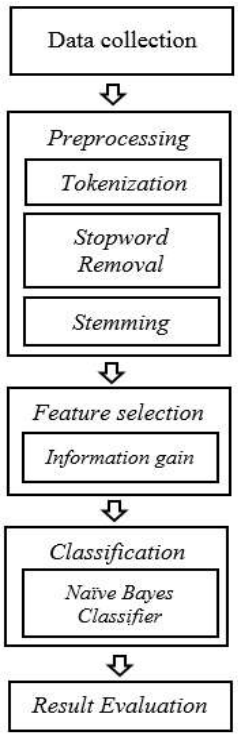


Fig. 1. Research Methodology

A. *Data Collection*

The data collection is conducted with tweet data crawling using the facility of Application Program Interface (API) which is provided by Twitter. Twitter Search API has weakness which can only search on the published tweets in the last seven days. The keyword popularity of the movie title used will affect the data potential and variation which can be crawled. In this research, it is chosen 12 movie titles manually based on the popularity in 2016. The movie titles are *Batman vs Superman*, *Suicide Squad*, *Jason Bourne*, *Sausage Party*, *Ouija*, *Swiss Army Man*, *Doctor Strange*, *Hacksaw Ridge*, *Passenger*, *Star Wars: Rogue One*, *Arrival*, and *La La Land*.

The next process, the crawling result is labeled manually into three classes which are positive, negative, and neutral. The crawling result sample and tweet labeling can be seen in the Table I.

TABLE I. SAMPLE OF THE CRAWLING RESULT AND CLASS LABELLING

Keyword	Result	Class
"suicide squad"	Suicide Squad is a great movie! It had a beast soundtrack too!	Positif
"batman vs superman"	Batman vs. superman was a disappointment .	Negatif
"jason bourne"	I just finished watching Jason Bourne.	Netral

B. Text Preprocessing

There are three steps in doing preprocessing process, which are Tokenization, Stopword Removal and Stemming[14].

1) *Tokenization.* Tokenization is a popular term for lexical analysis which is the stage of cutting input strings into phrases, words, symbols, or other elements called tokens [15]. In the tokenization process, all the input string data are changed into lowercase format first. Then, the character, number, and the punctuation but alphabet are deleted. The data, then, are cut being token which will be saved into database. The result sample of tokenization can be seen in the Table II:

TABLE II. SAMPLE OF TOKENIZATION RESULT

Text	Batman vs. superman was a disappointment.
Token	batman vs superman was a disappointment

2) *Stopword Removal.* The stopword removal process or filtering will select the words based on the database in the stopword list and wordlist. The stopword list data contain insignificant words in the classification process. Words that appear too often in a document can be considered insignificant because they have no statistical relevance so they can be eliminated from documents[16]. In this research, the movie titles are included into insignificant words, so it needs to go into the stopword list. The token, as the result of stopword list selection, then is matched to the English dictionary which has been verified in the wordlist database. In this research, we uses The Tournament Wordlist which became the standard of the scrabble game in America and contained 178.686 word databases.

3) *Stemming.* Stemming process aims at changing the token into base word or stem so that there is no word redundant in the database. The assumption used in the Stemming process is that each word is semantically related. The stemming process will significantly influence the language by having more complex word morphology [17]. Stemming method used in this research is Porter Stemmer. The stemming result samples can be seen in the Table III.

TABLE III. SAMPLE OF STEMMING RESULT

Derivational Word	**Stemming**
confus, confused, confusedly, confusion, confusions	confus
disappointed, disappointment, disappointments	disappoint

C. Feature Selection

Before going to the sentiment classification stage, the result feature of text preprocessing selected using one feature selection algorithm of filter type which is Information Gain. Information Gain is the most popular feature selection method in the text mining and frequently used as the text quality standard in the machine learning field [18]. The Information Gain method uses entropy to determine the best feature. The entropy states the uncertainty measurement probabilistically. The entropy value is firstly calculated by using equation (1) to obtain gain value [19].

$$\boldsymbol{I(S) = -\sum_{i=1}^{m} p_i log_2(p_i)} \quad (2)$$

Whilst, the entropy equation from each class is:

$$\boldsymbol{E(A) = -\sum_{i=1}^{m} I(S) \frac{s_{1i} + s_{2i} + \cdots + s_{mi}}{s}} \quad (3)$$

Entropy functions to show the variant similarity value on an attribute. The *gain* value is obtained from the total entropy subtracts attribute loyal entropy, thus it is obtained the calculation result in the form of variation value on each attribute. The *gain* value itself is calculated by using equation (4).

$$\boldsymbol{G(A) = I(S) - E(A)} \quad (4)$$

The result is that each keyword will gain certain gain value. Then, in the sentiment classification process, the researcher will test the gain value limit or threshold which is most appropriate in producing the best accuracy with an optimum run-time.

D. Data Classification

In this stage, it is calculated the word emergence probability on the data to determine a tweet as the member of positive, negative, or neutral class. The classification of the *Naïve Bayes Classifier* method is yielded by calculating likelihood each feature in the tweet with the implementation of the equation (7). Then, $P(TWEET|C_i)P(C_i)$is evaluated in every C_i class. The most calculation result shows the class prediction of the tweets.

E. Result Evaluation

The result evaluation in this research uses confusion matrix with the dimension of 3x3. The dimension is suited with the three tweet classes which are positive, negative, and neutral. Table IV shows the confusion matrix equation being used.

TABLE IV. THREE CONFUSION MATRIX CLASSES

		Predicted Class		
		C1	C2	C3
Actual Class	C1	*True Positives (TP)*	*False Positives (FP)*	*False Positives (FP)*
	C2	*False Negatives (FN)*	*True Negatives (TN)*	*False Negatives (FN)*
	C3	*False Netral (FNR)*	*False Netral (FNR)*	*True Netral (TNR)*

The accuracy calculation for this *confusion matrix* 3x3 is as follows :

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN+TNR}}{\boldsymbol{P+N+NR}} \times \mathbf{100\%} \quad (5)$$

IV. RESULT AND DISCUSSION

A. The Result of the Research Data Collection

The total data obtained and used in this test are 8,231 tweets. The crawled data are the tweets with English language. The time period of data collection and the total data based on the movie title can be seen in the Table V.

TABLE V. DATA COLLECTION RESULT

Keyword	Time	Total
Batman vs Superman	August 2016	715
Suicide Squad	August 2016	749
Jason Bourne	August 2016	462
Sausage Party	November 2016	555
Ouija	November 2016	113
Swiss Army Man	November 2016	131
Doctor Strange	November 2016	91
Hacksaw Ridge	November 2016	150
Star Wars : Rogue One	January 2016	93
Passengers	January 2016	127
Arrival	January 2016	103
La La Land	January 2016	121

The accuracy test level is conducted by selecting 3,410 tweets which divided into training dataset and test with the proportion of 70:30. As the result, 2,400 training dataset and 1,010 test dataset are selected for the accuracy test. The more detail dataset divisions can be seen in the Table VI.

TABLE VI. THE TOTAL OF THE DATASET BEING USED

Dataset	Positive	Negative	Neutral	Total
Training	800	800	800	2.400
Test	417	279	314	1.010

B. Feature Extraction

The beginning feature which is successfully obtained from the tokenization process is 3,978 words. After the implementation of the stopword removal and stemming with the Porter Stemmer algorithm, it is resulted 2,021 words. The final extraction sample from each tweet can be seen in the Table VII.

TABLE VII. THE EXTRACTION RESULT SAMPLE

Tweet Data	Feature	Class
Suicide Squad was actually really good, recommend v much	'actual', 'realli', 'good', 'recommend', 'much'	Positive
2 hours of my great life was wasted watching Jason Bourne.. #weepgang	'hour', 'great', 'life', 'wast', 'watch'	Negative
Finally watching Batman vs. Superman. I hope it makes me laugh as much as the first Superman did.	'final', 'watch', 'hope', 'make', 'laugh', 'much', 'first', 'did'	Neutral

C. *Gain Value Calculation*

The next process is gain calculation on each word. The accuracy ratio based on the gain value threshold will be conducted in the classification test. The average score of gain from this calculation is 0.008413. Table VIII shows 10 features which have the highest gain value or score.

TABLE VIII. HIGHEST GAIN RATIO FEATURE

No	Feature	Gain
1.	watch	0.398814
2.	just	0.165346
3.	good	0.134199
4.	see	0.134199
5.	like	0.117702
6.	you	0.108359
7.	have	0.0975346
8.	fuck	0.0860267
9.	not	0.073852
10.	final	0.0708243

D. *Data Classification Test*

In this stage, it is used the test data which are 1,010 tweets. After the preprocessing, there are 16 castaway tweets, so that the test data are reduced to 994 tweets. The feature which is successfully extracted from the test data then are calculated by using Naive Bayes Classifier algorithm as follows:

- The prior probability with the distribution of the same training data in each class yields:

$$P(C) = \frac{800}{2400} = 0,33333$$

- Next step is calculating the conditional probability or the likelihood with Laplacian correction. The sample of calculation result can be seen in the following test data:

La la Land was incredible last night

After the preprocessing, then it is obtained the features as below:

incred, last, night

Then, the class probability from the feature is calculated:

$$P(TWEET|C_{POSITIVE}) = \frac{6+1}{800+3} \times \frac{17+1}{800+3} \times \frac{11+1}{800+3} = 2,92015 \times 10^{6}$$

$$P(TWEET|C_{NEGATIVE}) = \frac{2+1}{800+3} \times \frac{15+1}{800+3} \times \frac{7+1}{800+3} = 7,41625 \times 10^{7}$$

$$P(TWEET|C_{NEUTRAL}) = \frac{0+1}{800+3} \times \frac{2+1}{800+3} \times \frac{8+1}{800+3} = 5,21455 \times 10^{8}$$

The result of the Likelihood calculation becomes the main parameter because the condition of the prior probability in each class has the same value or score. The class prediction of the above likelihood result is

$P(TWEET|C_{POSITIVE}) > P(TWEET|C_{NEGATIVE}) > P(TWEET|C_{NEUTRAL})$.

Thus, the class prediction of the tweet with the id 29 is positive.

The classification model in this research is yielded from the analysis towards the test data. The process to obtain the precise model is through the test of the test data accuracy of some gain value thresholds. The test result based on each threshold can be seen in the Table IX.

TABLE IX. THE CLASSIFICATION ACCURACY TEST

No	*Thres hold*	Total Feature	Correct Prediction	Incorrect Prediction	Accuracy
1.	0	2021	786	208	79,07%
2.	0.0055	1062	808	186	81,29%
3.	0.006	713	817	177	82,19%
4.	0.007	536	800	194	80,48%
5.	0.0075	445	804	190	80,88%
6.	0.008	376	795	199	79,98%
7.	0.009	322	781	213	78,57%
8.	0.01	285	768	226	77,26%

Table IX shows that the threshold gain value which is most optimum is 0.006. The threshold is able to reduce the total feature until 65% from the previous 21021 total keywords. It only needs 713 keywords to gain the best accuracy in this research. The classification result of the optimum threshold is evaluated by using confusion matrix with the dimension 3x3 written in the Table X.

TABLE X. THE EVALUATION OF THE CONFUSION MATRIX

n = 994		Predicted Class			
		Positive	Negative	Neutral	Class Precision
Actual Class	Positive	367	30	20	88,01%
	Negative	10	260	9	93,19%
	Neutral	53	55	190	63,76%
	Class Recall	85,35%	75,36%	86,76%	

The information which can be obtained from the table of the Confusion Matrix covers four parameters which are class recall, class precision, error rate, and accuracy. Table X has contained class recall and class precision information clearly.

The sentiment classification of the movie reviews with the combination method of information gain and Naïve Bayes Classifier in this research successfully gets the 82.19% accuracy level. Table 12 also shows the low of the precision of the neutral class. The total of the misclassification in the neutral class reaches 111 out of 317 test data. The number of this misclassification due to too many words of neutral class which are duplicate with words in the negative and positive class.

V. CONCLUSION

Based on the research finding and result which has been conducted, it can be drawn some conclusions that the classifier accuracy level reaches 82.19%. That accuracy is the highest accuracy with the gain value threshold 0.006. The class prediction with the highest precision is obtained by negative class with the percentage of 93.19. Neutral becomes the class with the lowest precision with only 63.76% as the result.

This research still has weakness that it has some misclassifications. The class precision in the neutral data becomes the main problem.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, Third Edit., vol. 3. Morgan Kaufmann, 2012.

[2] H. Rui, Y. Liu, and A. Whinston, "Whose and What Chatter Matters? The Effect of Tweets on Movie Sales," *Decis. Support Syst.*, vol. 55, no. 4, pp. 863–870, 2013.

[3] F. Morstatter, S. M. Ave, S. M. Ave, and K. M. Carley, "Is the Sample Good Enough ? Comparing Data from Twitter ' s Streaming API with Twitter ' s Firehose," *arXiv Prepr. arXiv1306.5204*, 2013.

[4] B. Liu, *Sentiment Analysis and Opinion Mining*, vol. 5, no. 1. San Rafael: Morgan & Claypool, 2012.

[5] R. P. Schumaker, A. T. Jarmoszko, and C. S. Labedz, "Predicting wins and spread in the Premier League using a sentiment analysis of twitter," *Decis. Support Syst.*, vol. 88, pp. 76–84, 2016.

[6] B. Pang and L. Lee, "Opinion Mining and Sentiment Analysis," *Found. Trends Inf. Retr.*, vol. 2, pp. 1–133, 2008.

[7] R. Feldman and J. Sanger, *The Text Mining Handbook*. New York: Cambridge University Press, 2007.

[8] Motion Picture Association of America (MPAA), "Theatrical Market Statistics 2015," pp. 1–28, 2015.

[9] J. R. Pentheny, "The Influence of Movie Reviews on Consumers," *Honor. Theses Capstones*, p. 265, 2015.

[10] S. M. Basheer and S. Farook, "Movie Review Classification and Feature based Summarization," *Int. J. Comput. Trends Technol.*, vol. 4, no. 7, pp. 2335–2339, 2013.

[11] V. Chandani and R. S. Wahono, "Komparasi Algoritma Klasifikasi Machine Learning Dan Feature Selection pada Analisis Sentimen Review Film," *J. Intell. Syst.*, vol. 1, no. 1, pp. 56–60, 2015.

[12] L. D. Utami and R. S. Wahono, "Integrasi metode information gain untuk seleksi fitur dan adaboost untuk mengurangi bias pada analisis sentimen review restoran menggunakan algoritma naive bayes," *J. Intell. Syst.*, vol. 1, no. 2, pp. 120–126, 2015.

[13] L. L. Dhande and P. G. K. Patnaik, "Analyzing Sentiment of Movie Review Data using Naive Bayes Neural Classifier," *Int. J. Emerg. Trends Technol. Comput. Sci.*, vol. 3, no. 4, pp. 313–320, 2014.

[14] A. A. Kumar and S. Chandrasekhar, "Text Data Pre-processing and Dimensionality Reduction Techniques for Document Clustering," *Int. J. Eng. Res. Technol.*, vol. 1, no. 5, pp. 1–6, 2012.

[15] T. Verma and D. Gaur, "Tokenization and Filtering Process in RapidMiner," *Int. J. Appl. Inf. Syst.*, vol. 7, no. 2, pp. 16–18, 2014.

[16] A. Hotho, A. Nürnberger, and G. Paaß, "A Brief Survey of Text Mining," *LDV Forum - Gld. J. Comput. Linguist. Lang. Technol.*, vol. 20, pp. 19–62, 2005.

[17] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing Techniques for Text Mining - An Overview," *Int. J. Comput. Sci. Commun. Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[18] Y. Yang and J. O. Pedersen, "A Comparative Study on Feature Selection in Text Categorization," *roceedings Fourteenth Int. Conf. Mach. Learn.*, vol. 20, no. 15, pp. 2429–2437, 1997.

[19] A. G. Karegowda, A. S. Manjunath, and M. A. Jayaram, "Comparative Study of Attribute Selection Using Gain Ratio and Correlation Based Feature Selection," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 271–277, 2010.

Interaction Design Consideration for Senior High School Students: A Usability Evaluation of Go-Jek Mobile Applications

Rani Laple Satria Putra
Magister Teknik Informatika
Universitas Atma Jaya Yogyakarta
Yogyakarta, Indonesia
rlsp1990@gmail.com

Joanna Cristy Patty
Magister Teknik Informatika
Universitas Atma Jaya Yogyakarta
Yogyakarta, Indonesia
Cristyjoanna18@gmail.com

Djoko Budiyanto Setyohadi
Magister Teknik Informatika
Universitas Atma Jaya Yogyakarta
Yogyakarta,Indonesia
djoko.bdy@gmail.com

Abstract— Today Mobile applications have grown tremendously. Since it influences many aspects, therefore, the quality measurement of the mobile application is needed, and it requires a thorough system evaluation of the usefulness of the application. The purpose of this study is performed by a feasibility test which uses structured data analysis methods. Data collections are completed by distributing questionnaires to 53 students of SMA Negeri 1 Depok, Sleman, Yogyakarta, and the collected data were analyzed using Framework Analysis (FA) and Usability Problem Taxonomy (UPT) method. The analysis includes many processes i.e., arrange, classify, classify and severity measurement. The results of the study yielded 12 usability problems with the severity of 2.33 (medium). Go-Jek order view and Go-Life View have the highest severity with 4 (critical), Go-Jek Home View, Go-Pay View and Orders View with 3 (severe) severities, Help Menu View and Go-Jek View with severity 2 (severe). Most classification problems in UPT is related to Artifact Component. It can be categorized where main categories are Visualness (1), Language (7), and Manipulation (4.) Another component is Task Component which consist two main categories i.e. Task Mapping (5) and Task Facilitation (4). Our research shows that by using FA and UPT methods, the comprehensive evaluation can be performed.

Keywords: Evaluation Usability, Framework Analyses, UPT, Mobile Applications.

I. INTRODUCTION

Usability evaluation is one of the methods in Human-computer Interaction area which can be used to assess the quality of a system such as interface design, reliability, and security of a system[1]. Usability is essential, and it can be viewed as a critical factor in measuring the safety of a system primarily to achieve customer loyalty [2]. There are many characteristics of the user of applications. Furthermore the user characterstics will influence its usability aplications. An example the usability of one application for young people wil be different with the adult people. Shorly, the continous usability evaluation is needed exceedingly if the variety of the user is wide.

The previous research reported that the usability testing which is performed for measuring the upper school students regarding the need for usefulness measurement of its features [3]. Gojek is one of the popular mobile application in Indonesia. It is comfortable and user friendly.

However, it does not fit for some people such as high school student. For them, the apllication menus are not suitable, since not all of high school student are not familiar to use the application. Furthermore, the usability evaluation is needed ,and there are many method which is used by researchs regarding evaluating the usability that have been widely used such as Heuristic Evaluation [4], cognitive walkthrough [5] and Pluralistic walkthrough [6] are usability methods.

Unfortunately, there is a few researches on usability of Gojek Application, for high school students. This is due to the different characteristics of the users in the high school group with the users in general. Most of the studies that have been conducted only discuss the evaluation of an application on a method. The absence of a framework of a data collection analysis makes it difficult to do a systematic study. That's why a composite approach that is Framework Analysis [7] and Usability Problem Taxonomy is needed in fixing this problem [8]. Combining the two methods can be done in order to find and to evaluate the proper research so identification of the important issues can be performed.

II. LITERATURE REVIEW

A. Usability Mobile Applications

Usability testing based on ISO (International Standards Organization) usability is aimed to measure the effectiveness, efficiency and user satisfaction obtained for a particular purpose. Usability testing is performed by the proper user, the user who matches the criteria in evaluating usability [9]. Effectiveness can be obtained by how far the goal or task is received for the user to complete. Efficiency refers to an effort to achieve a particular purpose [10].

The recent review of the usability of mobile applications reported that it is done mainly using newer validating methods [11]. In this study, we argue that any usability testing method can produce an identification of different problems. According to Bevan, et al. that there are four perspectives affecting usability: (1) usability can be measured from product ergonomics (2) User psychology perspectives (3) User and

product interactions (4) Usability must be tailored to the characteristics of the users according to whom [12]

B. Framework Analyses

The FA method is a data analysis framework used during data collection as well as overall data analysis [13]. In doing the study, several steps must be done. That is, the election of data obtained, mapping and sorting data based on the grouping of existing problems. The data collected can be quantitative data. The FA method is used to analyze data descriptively and textually to produce valid quantitative data [14]. Framework Analysis Stages: (1) familiarization of data, (2) Identifying data, (3) Indexing data, (4) Charting Data (5) Data Mapping.

C. Usability Problem Taxonomy

UPT is one useful method to classify usability problems by type. The UPT contains an artifact component a and a task component [15] The affiliate component comprises three categories (Visualness, Language, and manipulation) and a Task-mapping and Task Facilitation task. Each major category contains several sub-categories (for example, the group of visualness includes five sub-categories: object (screen) layout, object appearance, object movement, presentation of information/result, non-massage feedback). Other categories also contain subcategories. To classify the given problem, an evaluator uses the artifact component and task component [15].

However, in a component, categories and subcategories at a certain level are equally proprietary, the classification in a given element produces a final categorization [16]. To classify the problem within a component, the evaluator takes place as deeply as possible in that component, that is, until no further progress can be made. The evaluator then classifies the problem as deep as possible in the other ingredients. The classification process along the component can yield one of three results: Full Classification (FC), Partial Classification (PC), and Null Classification (NC) [15]. The use of UPT is still very little in the research, but its classification method can provide an excellent description of the dimensions of the problem of detected usability. Can be concluded UPT is a technique that is less well known but very good in its use.

III. METHODOLOGY

A. Data Collection

Usability evaluation can be done with several different methods. Inspection methods such as heuristic evaluation are one of the usability methods widely used by experts to identify problems in usability. [17] [18]. In carrying out data collection method is done by questionnaire with quantitative data that is by doing a questionnaire on high school students who become respondents by filling in the choice of answers [19][20].

The first section consists of brief questions about students' experiences about smartphone experience, internet experience, and usage, their experiences, and perceptions about Gocek applications, and the services they often use in Go-Jek applications. The given questionnaire focused on respondents' background questions.

B. Study Procedure

The participants are students of SMA Negeri 1 Depok, Sleman, Yogyakarta. The criteria used for the usability research participant are three i.e.: (1) Senior high school student (2) Using and owning the smartphone (3) Accustomed to using the internet. Furthermore, students who have and have not used the Go-Jek application are included in the evaluation of Go-Jek Usability application evaluation. The accident sampling is performed and the data collection is done in the sourounding of SMA Negeri 1 Depok.

Students will be given questions about Go-Jek application experience. If the student has not used the Go-Jek app, then the student is assigned to handle the Go-Jek app. After that, the students make a transaction on the service menu found on the Go-Jek application. Students carry out the use of applications gradually and according to the procedures that are executed. Then after doing the application students are evaluated whether there are questions on the form. Evaluation begins with students filling out a short demographic questionnaire. Topics include age, gender, class, smartphone usage, internet usage and brief information about the Go-Jek app [21].

After performing the first task, the student continues the next work of filling out a usability evaluation questionnaire based on her experience while doing interaction with the application. This questionnaire contains questions about the evaluation of the GO-Jek application system. These questions serve as data to be analyzed at a later stage. The task completion stage is completed, and the students finish the assigned task and collect the completed questionnaires [22].

C. Data analyses

At this stage, we will identify the data obtained from the usability questionnaire. After that, the data has been obtained identified using computer software such as Microsoft Excel and Microsoft Word [23]. Analysis performed by FA and UPT.

The framework analysis is carried out during the analysis process of the data that has been collected. Then, to perform the data analysis, five steps are required in the FA method [16] ie (1) introduction of data, (2) identifying the theme and framework to be used (3) classify usability problems obtained using UPT, (4) arrange or map results on the Go-Jek mobile app (5) are planned and interpreted. For the scale of ratings using a level that is (1) low, (2) moderate, (3) serious, (4) critical [24]. From the scores, the value will be generated the average amount of each display as well as the overall view system.

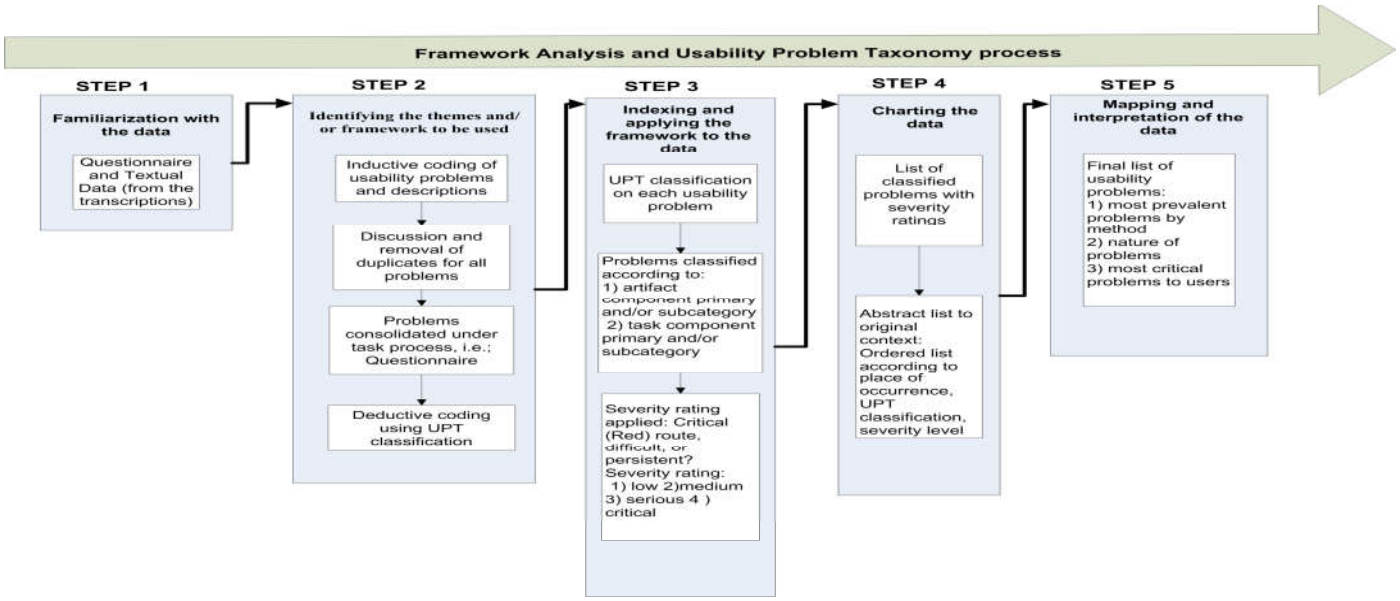


Fig 1. Application of the FA Method and UPT [16]

IV. Result and Discussion

A sample of 53 senior high school students consisting of 16 men and 37 women. Mostly aged 15-17 years with a range of age 14-17 years. The majority of respondents are in class X and XI, and most live with parents either with mom or dad. Ninety-six percent have used Go-Jek mobile apps.

The sample is a regular smartphone user. For example, 100% use your smartphone every day. 39.62% using internet 3 hours to five hours, 26.41% use internet more than five lam on a school day and 71.69% use internet with the data packet. The majority agree that the initial appearance of mobile apps is attractive, and the majority of respondents use Go-Ride, Go-Food and Go-Car services a lot.

A. Usability Problem

There are 34 usability problems proposed to 53 students on data collection methods. 34 such problems were carried out with a questionnaire. The issues that have been described are grouped into 12 usability issues. The average severity is 2.33 or medium. The results are based on the most critical issues listed on the scene and shown at table 1 below.

According to Table 1, the highest usability problem and severity are in Go-Jek Home View which has three usability problems with three severity, Go-Pay View with two issues with seriousness of three, Order's View Menu with two points with severity three, Help Menu Problem usability and the highest severity is on Go-Jek Home View which has three usability issues with three severity levels, Go-Pay View with two problems with severity of three, Order's Menu View with two points with severity three, Help Menu View with two issues and the severity of the two, Go-Jek View one issue of seriousness two, Go-Jek Order View with one problem with four severity, Go-Life View with one question with four rigor.

Based on the data Go-Jek Home View has the highest level of problems, but at Go-Jek Order View and Go-Life View have the highest severity level although each has one problem. Thus the severity and mobile usability issues of Go-Jek applications exist in high school students focused on Go-Jek Home View, Go-Jek Order View, and Go-Life Order View. Graph of severity and usability problem can be seen in fig. 2

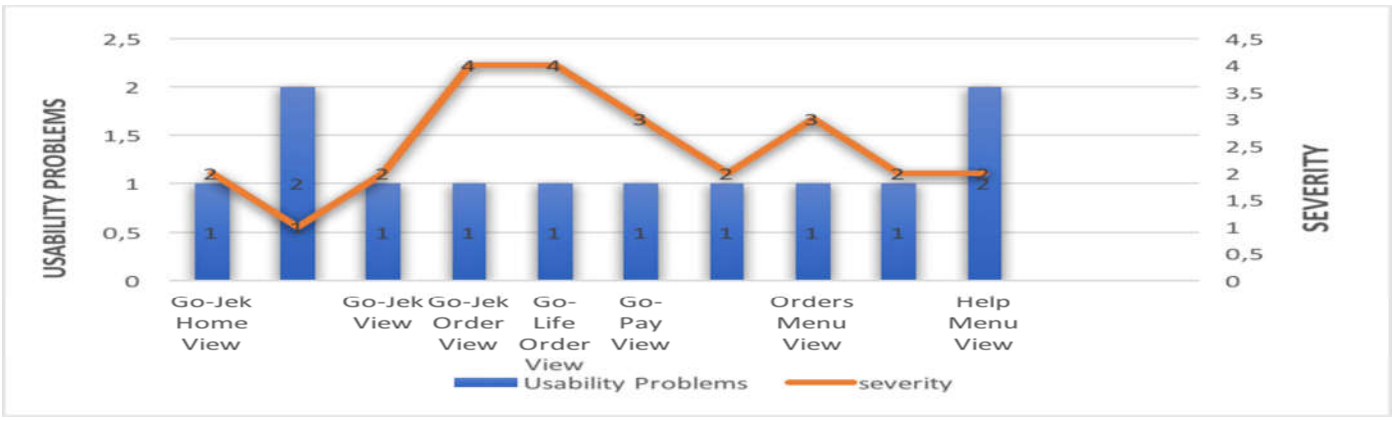


Fig. 2 Usability Problems and Severity Rating

TABLE I. USABILITY PPROBLEM CLASSIFICATION

Description	Place of Occurrence	% of pat. detecting method	UPT Classification		Severity
			Artifact	Task	
Difficult to understand word and icon on Go-Jek main menu	Go-Jek Home View	17%	Language-Naming/Labeling (FC)	NC	1
Difficult to understand application's language	Go-Jek Home View	11,32%	Language-Naming/Labeling (FC)	NC	1
No feedback from application after inputting any action	Go-jerk View	49%	Language-Order Wording/Feedback Messages (FC)	task facilitation-task/function automatic (FC)	2
Application's interface is inconsistent and confusing	Go-Jek Home View	38%	visualness-object(screen)layout (FC)	NC	2
Difficult to make the order on the Go-Jek application	Go-Jek order View (Go-Send, Go-Mart, Go-Shop, Go-BlueBird, Go-Point, Go-Pulsa, Go-Bills, Go-Box, Go-Med, Go-Pay)	78%	Manipulation-Cognitive Aspect (FC)	task mapping-navigation (FC)	4
Difficult to make an order on Go-Life application	Go-Life Order View (Go-Massage, Go-Auto, Go-Clean, Go-Glam)	90,10%	Manipulation-Cognitive Aspect (FC)	task mapping-navigation (FC)	4
Difficult To Understand How To Top-Up Go-Pay Balance	Go-Pay View	51%	Manipulation-Cognitive Aspect-Direct Manipulation (FC)	Task Mapping-Navigation (FC)	3
Difficult To Understand How To Check The Go-Pay Balance	Go-Pay View	43%	Manipulation-Cognitive Aspect-Direct Manipulation (FC)	Task Mapping-Interaction (FC)	2
Difficult To Check The Status Of The Order	Orders Menu View	28%	Language-Order Wording/Feedback Messages (FC)	Task Facilitation-Task/Function Automatic (FC)	2
Difficult To Understand About Given Help And Documentation	Help Menu View	43,40%	Language-Order Wording/Feedback Messages (FC)	Task Mapping-Navigation (FC)	2
Difficult To Understand The Reason For The Fail Transaction	Orders Menu View	64%	Language-Order Wording/Feedback Messages (FC)	Task Facilitation-Keeping The User Task On Track (FC)	3
Difficult To Understand Help Desk Information	Help Menu View	49%	Language-Order Wording/Other System Messages (FC)	Task Facilitation-Task/Function Automatic (FC)	2

The most critical usability problem or severity level 4 is found in Go-Jek Order View and Go-Life Order View. In this menu, there are service menu services for household purposes such as massage, salon, workshop, and others. On this menu, students are tasked to place orders on services provided in Go-Jek Order View and Go-Life Order View. To do the students have difficulty in ordering sequentially because many commands and options make students become confused in order. On this issue are grouped into (Artifact) Manipulation-Cognitive Aspect. Both of these problems are usability problems with the highest severity in high school students obtained from questionnaire data collection.

On the next problem is the severity level 3 is on Go-Pay View and Orders Menu View. The problem with Go-Pay View is the difficulty for students to add balance due to the many steps that must be done and difficult to understand, thus making the students confused in making new balances. In Orders Menu View there are difficulties when the transaction fails. At the time of the operation, there is a difficulty in the students because challenging to understand the command for the next step.

Other difficulties are also available in different menus, Go-Jek Home View, Go-Jek View and Help Menu View. On these three menus, the severity level is only at level 1 and 2. Only a small percentage of students have difficulty making use of this menu. From the results obtained the majority of problems are in trouble understanding in doing feedback after accessing the list. So it takes instructions in this menu to facilitate students at the time of doing feedback.

B. Results for Methodological Approach

Data collection techniques with questionnaires that have been done to discuss 34 problems in the form of questions on the survey. From this questionnaire can facilitate the respondent in answering questions by the issue of usability. The inquiry is very appropriate in identifying problems especially for high school students because it is easily understood in solving the problem. After the data is obtained, the grouping of the issues can be performed and analyzed quickly.

C. Classification of Problems

This figure shows the distribution of problems from each UPT classification class and its sub-category. Usability problems are grouped in the artifact component of Visualness, Object (screen) layout, Manipulation, Cognitive aspect, Direct Manipulation, Language, Naming/labeling, Other Wording, Feedback Massage, Other System Massage and Manipulation, Cognitive Aspect, Direct Manipulation. Next on Task Component problems are Task Mapping, Interaction, Navigation and Task Facilitation, Task / Function Automation, Keeping user task on track. For problem distribution, we classify all issues to the component artifact and component tasks. There are three component tasks of the results of this study that can not be classified. In the Artifact Component, Language receives the highest volume of full classification with 7 Full Classification. While in Task Component, Task Mapping is the highest volume.

In fig.3 can be a delight, classification problem of the whole system. Visualness has one problem: Object (screen) layout. Language has seven issues: 2 difficulties on Naming / Labeling, five issues with Other Wording, four points on Feedback Massage and one problem with Other System Text. Manipulation has four problems namely four issues on Cognitive aspect with two questions in Direct Manipulation subcategory. Other issues are also present in Task Mapping with five items that are one problem in Interaction and four problems on Navigation. Task Facilitation with four issues namely Task function three problem and Keeping the user task on track one problem. From this problem shows that most of the usability problems occur in the design and structure of the system cannot accommodate, especially in high school students.



Fig.3 Artifact and Task Component Classifications

D. Usability problem classifications

There are 34 usability problems identified by the questionnaire method, then classified to be grouped into 12 usability problems. The majority of the issues are Go-Jek Home View with severity 2 and 1. The next point is the Help Menu View with severity 2, Go-Pay Menu View with severity 3 and 2 and Orders view menu with severity 3 and 2. The next problem is on Go -Jek View with severity 2. Last is Go-Jek Order View and Go-Life Order View which have the least problem but have the highest severity of value 4 (critical). Problems that occur in these two views are student's difficulty in making transactions, make orders and difficulty in following the commands that exist on the menu.

Based on UPT classification, most of the problems are in Language-Other Wording. The next question is in Manipulation; Cognitive Aspect also dominates from usability problem so that Language and Cognitive Aspect problem becomes the main problem in the content of this system. The results obtained are the difficulty of students in navigating, especially in order and transaction, so the need for a solution to the problem so that the system can be easily used by high school students.

E. Structured data analysis

In compiling data collection methods, the FA method is very useful in encoding most of the existing data [16]. However, in this FA method, we make modifications that are adapted to the UPT in classifying usability issues. Thus the FA method is advantageous in modifying the evaluation of usability by the existing problems.

UPT is very useful in describing the problem. By using UPT then the problem can be classified into Full, Partial and Null Classification [16]. In this study, only a few usability

problems that can be classified into the category of artifact component that is Visualness, Naming / Labeling, Language, but some issues have been classified not into the type of Task Component so cannot describe more details of the problem.

In this study, we found that the classification of each problem with the UPT is significant so that it can determine the more detailed and hierarchical. In this study, we also found that the problem can be well explained so that we can easily categorize the problem according to the nature of the same problem [17]. Also, we can find that determining severity can help in solving problems so that it can efficiently analyze issues and approach better solutions for the next.

V. CONCLUSIONS

From the results of this study shows that in the mobile application Go-Jek in usability test shows there is two menu that has a high rating severy rating with a value of 4 (critical) that is on Go-Jek Order View and Go-Life Order View. In the result, 78% of respondents had difficulty in the Go-Jek Order View menu and 90.10% in Go-Life Order View. From these results indicate that it is necessary to adjust the interface design and language, especially the commands in making transactions so easy to use by users in high school students.

In the future research usability evaluation is indispensable in evaluating interface design, especially in the use of user interaction with mobile applications so that it can adjust the interface design with user characteristics.

In this study also proved that the FA method could also be used in analyzing quantitative data so that the resulting data can be grouped in a structured and systematic. With the FA method that can guide researchers in analyzing data and UPT is very useful as a more conferencing classification scheme in determining the severity of user problems. In this case, it also helps to categorize and identify the kind of the issues that can be useful for the designer.

REFERENCES

[1] J. A. Sheikh and A. Arshad, "Advances in Usability and User Experience," vol. 607, 2018.

[2] N. Ahmad, A. Rextin, and U. E. Kulsoom, "Perspectives on usability guidelines for smartphone applications: An empirical investigation and systematic literature review," *Inf. Softw. Technol.*, vol. 94, pp. 130–149, 2018.

[3] T. Kajiyama, "An improved botanical search application for middle- and high-school students," *Educ. Inf. Technol.*, vol. 21, no. 6, pp. 1821–1836, 2016.

[4] I. Almarashdeh and M. Alsmadi, "Heuristic evaluation of mobile government portal services: An experts' review," *2016 11th Int. Conf. Internet Technol. Secur. Trans. ICITST 2016*, pp. 427–431, 2017.

[5] A. Hussain and E. O. C. Mkpojiogu, "Usability evaluation techniques in mobile commerce applications: A systematic review," no. August, p. 020049, 2016.

[6] T. Hollingsed and D. G. Novick, "Usability inspection methods after 15 years of research and practice," *Proc. 25th Annu. ACM Int. Conf. Des. Commun. - SIGDOC '07*, no. October 2007, p. 249, 2007.

[7] N. K. Gale, G. Heath, E. Cameron, S. Rashid, and S. Redwood, "Using the framework method for the analysis of qualitative data in multi-disciplinary health research TL - 13," *BMC Med. Res. Methodol.*, vol. 13 VN-r, no. 1, p. 117, 2013.

[8] R. Geng, M. Chen, and J. Tian, "In-process usability problem classification, analysis and improvement," *Proc. - Int. Conf. Qual. Softw.*, pp. 240–245, 2014.

[9] A. Widyanti, S. Amalia, Q. Ainizzamani, and A. Usability, "Usability Evaluation of Online Transportation ' User Interface," pp. 82–86, 2017.

[10] Z. Huang and M. Benyoucef, "Usability and credibility of e-government websites," *Gov. Inf. Q.*, vol. 31, no. 4, pp. 584–595, 2014.

[11] M. W. Iqbal, N. Ahmad, and S. K. Shahzad, "Usability evaluation of adaptive features in smartphones," *Procedia Comput. Sci.*, vol. 112, pp. 2185–2194, 2017.

[12] R. Inostroza, C. Rusu, S. Roncagliolo, V. Rusu, and C. A. Collazos, "Developing SMASH: A set of SMArtphone's uSability Heuristics," *Comput. Stand. Interfaces*, vol. 43, pp. 40–52, 2016.

[13] J. Ritchie and J. Lewis, "Qualitative Research Practice: A Guide for Social Science Students and Researchers," *Qual. Res.*, p. 356, 2014.

[14] A. Srivastava and S. B. Thomson, "Framework Analysis: A Qualitative Methodology for Applied Policy Research," *J. Adm. Gov.*, vol. 4, no. 2, 2009.

[15] S. L. Keenan, H. R. Hartson, D. G. Kafura, and R. S. Schulman, "Usability problem taxonomy: A framework for classification and analysis," *Empir. Softw. Eng.*, vol. 4, no. 1, pp. 71–104, 1999.

[16] M. Georgsson and N. Staggers, "An evaluation of patients ' experienced usability of a diabetes mHealth system using a multi-method approach," *J. Biomed. Inform.*, vol. 59, pp. 115–129, 2016.

[17] D. Quiñones and C. Rusu, "How to develop usability heuristics: A systematic literature review," *Comput. Stand. Interfaces*, vol. 53, pp. 89–122, 2017.

[18] R. Khajouei, S. Hajesmaeel Gohari, and M. Mirzaee, "Comparison of two heuristic evaluation methods for evaluating the usability of health information systems," *J. Biomed. Inform.*, vol. 80, pp. 37–42, 2018.

[19] J. Kaipio *et al.*, "Usability problems do not heal by themselves: National survey on physicians' experiences with EHRs in Finland," *Int. J. Med. Inform.*, vol. 97, pp. 266–281, 2017.

[20] D. B. Setyohadi, S. Kusrohmaniah, E. Christian, and L. Triani, "M-Learning Interface Design Based on Emotional Aspect Analysis M-Learnıng Interface Desıgn Based On Emotıonal Aspect Analysıs," no. January, 2017.

[21] S. Roy, P. K. Pattnaik, and R. Mall, "A quantitative approach to evaluate usability of academic websites based on human perception," *Egypt. Informatics J.*, vol. 15, no. 3, pp. 159–167, 2014.

[22] B. Brown, P. Balatsoukas, R. Williams, M. Sperrin, and I. Buchan, "Multi-method laboratory user evaluation of an actionable clinical performance information system: Implications for usability and patient safety," *J. Biomed. Inform.*, vol. 77, no. November 2017, pp. 62–80, 2018.

[23] D. J. Ward, C. Furber, S. Tierney, and V. Swallow, "Using Framework Analysis in nursing research: A worked example," *J. Adv. Nurs.*, vol. 69, no. 11, pp. 2423–2431, 2013.

[24] David Travis, "How to prioritise usability problems," 2009. [Online]. Available: https://www.userfocus.co.uk/articles/prioritise.html. [Accessed: 19-May-2018].

Prediction of Movie Sentiment based on Reviews and Score on Rotten Tomatoes using SentiWordnet

Suhariyanto
Faculty of computer science
Dian Nuswantoro University
Semarang, Indonesia
suhariyanto@dsn.dinus.ac.id

Ari Firmanto
Department of Informatics
Institute Technology Sepuluh
Nopember Surabaya, Indonesia
arifirmanto.17051@mhs.its.ac.id

Riyanarto Sarno
Department of Informatics
Institute Technology Sepuluh
Nopember Surabaya, Indonesia
riyanarto@if.its.ac.id

***Abstract*— With the number of films released each year, the movie review website is becoming more popular. One of the most referenced movie review websites is Rotten Tomatoes. Rotten Tomatoes recommend films based on their Tomatometer. Tomatometer represents the percentage of professional critic reviews that are positive or negative for a given film or television show. While fresh reviews represent positive sentiment, rotten reviews mean that the movie critics give the movie negative sentiments. Unfortunately, the method to determine the given score is not available to the public. Thus, the public does not know which parameter affect the prediction of the sentiment. This paper proposes a new method to predict the sentiment of the movie on the rotten tomatoes by combining the sentiment score from SentiWordnet and expert original score. the result of the experiment shows that the proposed method gives better F measure compared to those of the other methods with the value of 0.97.**

keywords—rotten tomatoes, sentiment analysis, sentiwordnet

I. INTRODUCTION

In the era of information technology, any information can be gained quickly such as movie information. Every year a new movie is released, it will be better if there is a recommendation to filter movie that people wants to watch. Nowadays, there are a lot of movie recommendations, one of them is rotten tomatoes. It gives the recommendation based on the score that calculates by the system using expert critic review. However, the process behind it is not publicly available.

Sentiment analysis is one of the methods to analyses user opinion from the review. From the previous research, many researchers using two approaches in sentiment analysis: lexical based approach and supervised machine learning approach. The Lexical approach using the dictionary that has sentiment polarity on them, such as SentiWordnet. Supervised machine learning approach using statistical analysis to predict sentiment polarity using the defined label as a target. The problem of supervised machine learning is dependent on the domain. If data set in the different domain then, machine learning need to be retrained to get high accuracy [1].

SentiWordnet [2] is a dictionary that has sentiment polarity based on Wordnet. Wordnet structure is a tree-like, that start with root word as hypernym and have branched from the root that called hyponym. It uses a random walk algorithm to give a score from Wordnet. To choose correct sense form synset (synonym set), word sense disambiguation algorithm is needed, because wordnet structure has the different meaning for the same word.

In RottenTomatoes, sentiment polarity is classified into two classes, rotten and fresh. Rotten is a negative sentiment and fresh is a positive sentiment. An expert does not only give their review, but they also give their rating called original score. The proposed method of this paper is to combined both expert review and expert original score. SentiWordnet used to extract sentiment score from review. To combined with the original score, normalization is needed for both values. The result of this is a feature that will be proceeded by logistic regression as a classification algorithm to predict movie sentiment.

II. RELATED WORK

Pang and Lee [3] used Naive Bayes, Maximum Entropy (ME), and SVM to classify binary sentiment on movie reviews. They collected the dataset from IMDB and used various features which were highest results come from SVM with unigram feature and got accuracy about 82.9%.

Dang et al [4] classified sentiment using SVM for a multi-domain dataset. They collected 305 positive reviews and 307 negative reviews from the digital camera. Information gain was used as a feature selection and got accuracy for 84.15%.

Agarwal et al. [5] classified sentiment from tweets using SVM (Support Vector Machine). They experimented with 11875 manually labeled tweets annotation. The authors used the combination of features which were unigram, sentiment score, and dependency tree. The highest accuracy of the model is 75.39%.

Xia et al. [6] used ensemble machine learning model to classified sentiment. The classifiers were Naive Bayes, Maximum Entropy, and SVM. The features that used was POS Tagging and word relation. Ensemble classifier gave the highest accuracy about 87.7%.

Pamungkas and Putri [7] used SentiWordnet to classified play store review dataset. They used first sense to choose synset from wordnet. The result of their experiment is about of 67.93%. Fika et al [8] used corpus-based emotion (CBE) to predict music emotion based on lyrics. CBE is the merging of two computational model which are WNA (Wordnet Affect Emotion) and ANEW (Affective Norms for English Words) using terms similarity, distancing of the node, and topic modeling [9].

Bagus et al [10] extracted aspect from product review and determined sentiment of each aspect. The Dataset used was product review from Amazon. The authors handled both local and global context from the review to improve sentiment analysis. In the field of business model, Endang et

al [11] built a lexical database using the same method with BabelNet for business process, it called B-BabelNet. BabelNet [12] is the lexical database of semantic network based on Wikipedia and Wordnet. B-BabelNet calculated semantic similarity in the domain of business process and it r caould be used to improve sentiment analysis in the business process.

III. RESEARCH METHOD

In this paper, the method is divided into crawling the data from Rotten Tomatoes, text preprocessing, feature extraction and sentiment classification. Detail processes can be seen in Figure 1, which is described as follows.

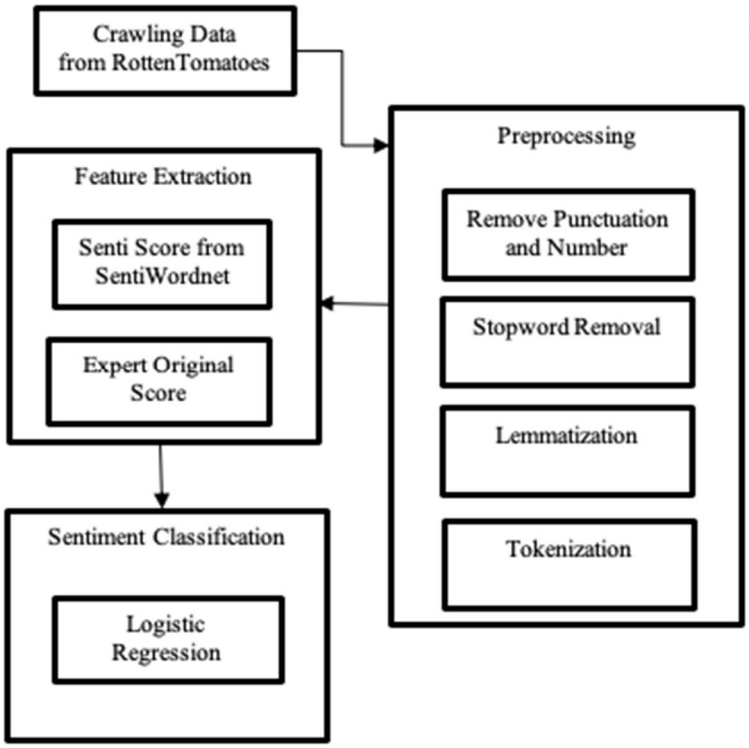


Figure 1. Propose Framework

A. Dataset Collection

Data is collected by crawling on rotten tomatoes. It should have an expert original score if it does not have then data is skipped. Then data is saved into CSV file. The total amount of data is 300, which were evenly selected with 150 data for rotten class and 150 data for fresh class. This balanced selection was carried out on purpose to get better result.

The purpose of the preprocessing is to prepare data and to remove noise from the data in order to increase of the accuracy for machine learning to learn pattern from the data. The steps are:

- **Remove Punctuation and Number** is the step to remove punctuation and number from the movie reviews.
- **Normalize into Lower Case** is the step to change all word into lower case.
- **Stop Word Removal** is the step to remove English stop word from text review. NLTK library is used for stop word removal.
- **Lemmatization** is the step to change a word into a lemma using the dictionary. In this paper, NLTK library used to change a word into a lemma.
- **Tokenization** is the step to change text review into the token of words.

C. Feature Extraction

The Feature that used for feature extraction is a sentiment score from SentiWordnet and expert original score. The detail process for each feature is:

- SentiWordnet

Score calculated using SentiWordnet. In this paper, only an adjective score and verb score used. SentiWordnet modeled part of speech (POS) from the treebank parser. Thus, score collected only if that word has POS tagging JJ, JJR, JJS, VB, VBD, VBG, VBN, VBP, and VBZ. The detail treebank parsers can be seen in Table 1.

TABLE I. TREE BANK PARSER

Tree Bank Parser	Definition
JJ	adjective
JJR	adjective, comparative
JJS	adjective, superlative
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non 3rd person singular present
VBZ	Verb, 3rd person singular present

To find the correct synset from SentiWordnet, adapted weighted graph [13] and lesk algorithm [14] are used. Sentiment score is calculated by the sum of the positive or negative score divided by the total of the positive and negative score. The equation can be seen in Equation 1, 2, 3, and 4.

$$pos_score = \sum_{i=i}^{n} pos_score_senti(i) \quad (1)$$

$$neg_score = \sum_{i=i}^{n} neg_score_senti(i) \quad (2)$$

$$total_score = pos_score + neg_score \quad (3)$$

$$senti_score = \begin{cases} \frac{pos_score}{tota_score} if\ pos > neg \\ \frac{neg_score}{total_score}\ if\ neg > pos \end{cases} \quad (4)$$

In the equation above, n is the total word from text review, pos_score_senti is individual word score from SentiWordnet, pos_score is the total positive score, neg_score_senti is individual word score from SentiWordnet, neg score is the total negative score, and senti_score is the final score used as a feature.

- Expert Original Score

For the expert original score, the value divided into categorical and numeric value because some expert use alphabet to give a score in their review. For categorical, the self-define rule used for mapping categorical into numeric value. The rule can be seen in Table 2

TABLE II. NORMALIZED CATEGORICAL SCORE

Expert Score	Transform Score
A	5
B	4
C	3
D	2
E	1
+	added by 0.25
-	subtract by 0.25

For the numerical score, must be normalized first because it contained difference range of value. The equation can be seen in Equation 5:

$$rating = 5 / max_org_rating * org_rating \quad (5)$$

In the Equation 5, max_org_rating is the maximum range rating from the expert original score and org_rating is the original score from the expert. After rating has obtained, then it will be divided by 5 which are the maximum rating score. This equation can be seen in Equation 6.

$$exp_final_rating = rating / 5 \quad (6)$$

D. Sentiment Classification

For sentiment classification, data will be modelled into the regression, where the goal is to predict the weight between sentiment score and expert original score. The weight implies the important features between them. Logistic regression is used in this paper, because the target is binary classification. The equation for logistic regression can be seen in Equation 7:

$$log_y = \frac{exp(\alpha * X1 + \beta * X2 + \gamma)}{1 + exp(\alpha * X1 + \beta * X2 + \gamma)} \quad (7)$$

Where γ is the bias, α and β is the weight, and X1, X2 is the features. To find the α and β that satisfy Equation 7 that minimizes the error, stochastic gradient descent used. The stochastic gradient descent is the variant of gradient descent that used only randomly one piece of the data from dataset in the calculation. The equation for stochastic gradient descent can be seen in Equation 8 and 9.

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y^i log_y + (1 - y^i) log(1 - log_y)] \quad (8)$$

$$\theta_j = \theta_j - \alpha J(\theta) \quad (9)$$

m is the total sample size, y is the ground truth of the rating data, and α is the learning rate.

IV. RESULT

To evaluate the results, the proposed method was compared with the baseline method. The baseline method that used in this paper is SentiWordnet, expert original score, and SVM (Support Vector Machine). SentiWordnet sentiment orientation calculated using Equation 1 and 2. If the positive score greater than the negative score than it classified as fresh otherwise is rotten. For the expert original score, to determined sentiment orientation threshold is used. The threshold value is 3 if the original score greater than or equal 3 then classified as fresh otherwise is rotten. And SVM using only Term Frequency-Inverse Document Frequency (TF-IDF) features with linear kernel. For evaluation, the confusion matrix will be used. The equation can be seen in Equation 10, 11 and 12:

$$precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (10)$$

$$recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (11)$$

$$F - measure = 2 * \frac{precision * recall}{precision + recall} \quad (12)$$

For SentiWordnet and expert original score, evaluation matrix will use all data because it does not have a model that needs to be trained. Otherwise, for the proposed method and SVM data will be split into training and testing to overcome overfitting on the data. Distribution of the data can be seen in Figure 2. From Figure 2, the total of rotten and fresh from SentiWordnet and Expert Original Score is 150. For SVM and the proposed method, the total of data is 29 for rotten and 31 for fresh. This is because it is using only testing data for the evaluation matrix. Confusion matrix comparation can be seen in Figure 3 and Table 3.

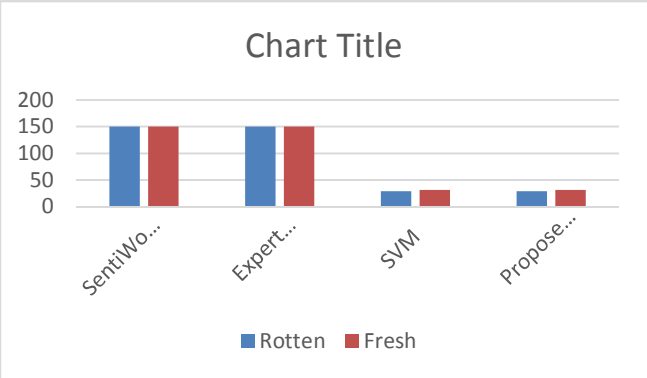


Figure 2. Data Distribution of Each Method

TABLE III. AVERAGE PRECISION, RECALL, AND F-MEASURE COMPARATION

Method	Precision	Recall	F Measure
SentiWordnet	0.51	0.51	0.51
Expert Original Score	0.92	0.91	0.91
SVM - TFIDF	0.77	0.77	0.77
Proposed Method	0.97	0.97	0.97

TABLE IV. PRECISION, RECALL, AND F-MEASURE COMPARATION OF FRESH SENTIMENT

Method	Precision	Recall	F Measure
SentiWordnet	0.51	0.52	0.51
Expert Original Score	0.86	0.98	0.92
SVM - TFIDF	0.74	0.79	0.77
Proposed Method	1.0	0.93	0.96

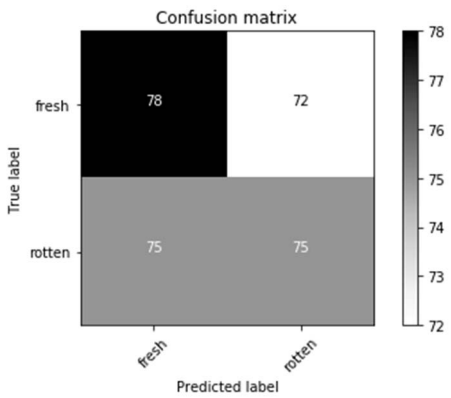
TABLE V. PRECISION, RECALL, AND F-MEASURE COMPARATION OF ROTTEN SENTIMENT

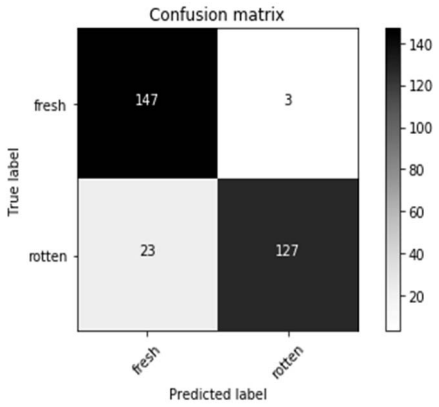
Method	Precision	Recall	F Measure
SentiWordnet	0.51	0.50	0.51
Expert Original Score	0.98	0.85	0.91
SVM - TFIDF	0.79	0.74	0.77
Proposed Method	0.94	1.0	0.97

From Table 3, our proposed method outperformed other methods with F1 measure 0.97. For SVM, using TF-IDF features depends on the word distribution. If the expert using various of words to describe their review, then data matrix will be sparse and good accuracy will be difficult to achieve. For the proposed method, using the combination of SentiWordnet and expert original score show good result. SentiWordnet cannot handle the review that has implicit sentiment, for instance in review: "*Maybe the quips and the punch-ups are there because they have to be, what with the film's steady parade of failure and even death. Plans Fail. Character fails. Even sacrifices fail. It's not exactly refreshing, but it is bracing, and even gratifying*". In that review, there are a lot of words that fail indicated to negative sentiment, for example is *failure*, *fail*, and *death*. Thus, using SentiWordnet will be classified as rotten but the ground truth is fresh. The expert original score is the second highest result from the baseline, this is because original score is the feature importance. Table 6 shows the importance of features from the trained model, and the negative value means that the corresponding feature pushes the classification more towards the negative class in this case, is rotten. Thus, the recall of the rotten class is higher than fresh class, this can be seen in Table 4 and Table 5. Thus, using only expert original score baseline good result can be achieved. But, the expert original score has lower precision than the proposed method. This can be seen from review "*Ultimately, it is more interesting to think about than it is to watch*" the expert gives rating 2 from 5, then it will be classified into rotten, but the ground truth is fresh. Our proposed method fixes the problem above and gets the closest result with rotten tomatoes ground truth.

TABLE VI. FEATURE IMPORTANCES OF LOGISTIC REGRESSION MODEL

	Sentiment Score	Expert Original Score
Coefficient	0.289	-2.313





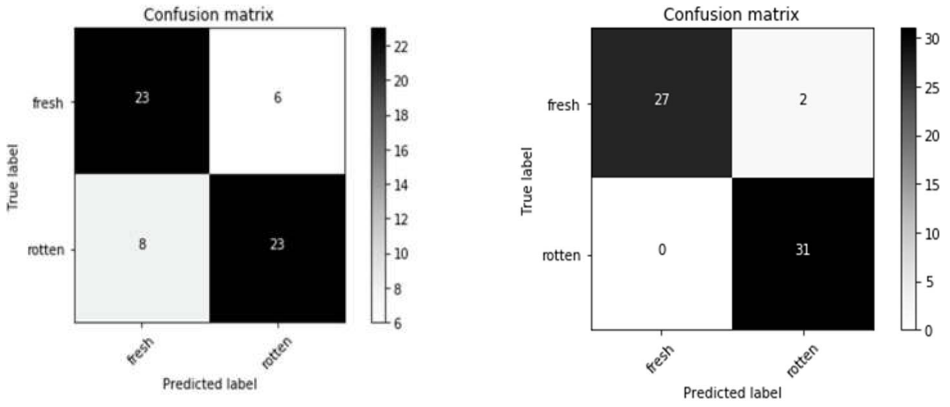


Figure 3. Top left is the Confusion Metrix of SentiWordnet, top right is the Confusion Metrix of Expert Original Score, bottom left is the Confusion Metrix of SVM using TF-IDF feature, and bottom right is the Confusion Matrix of the proposed method.

V. Conclusion and Future Work

The experiment showed that there were many implicit reviews from the expert that SentiWordnet had failed to recognize it, whereas the proposed method could recognize implicit reviews because of the additional features; i.e., sentiment score and expert original score.

For future work, we will implement imbalance algorithm to handle imbalance class. Also, the effect of incomplete data of expert original score will be analyzed.

References

[1] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, vol. 5, no. 1, pp. 1–167, 2012.

[2] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0 : An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining SentiWordNet," *Analysis*, vol. 10, pp. 1–12, 2010.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "{T}humbs up? {S}entiment classification using machine learning techniques," in *Proc. of the EMNLP'02*, 2002.

[4] Y. Dang, Y. Zhang, and H. Chen, "A lexicon-enhanced method for sentiment classification: An experiment on online product reviews," in *IEEE Intelligent Systems*, 2010, vol. 25, no. 4, pp. 46–53.

[5] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau, "Sentiment analysis of Twitter data," *Association for Computational Linguistics*, pp. 30–38, 2011.

[6] R. Xia, C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification," *Information Sciences*, vol. 181, no. 6, pp. 1138–1152, 2011.

[7] E. W. Pamungkas and D. G. P. Putri, "An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia," in *Proceedings - 2016 6th International Annual Engineering Seminar, InAES 2016*, 2017, pp. 28–31.

[8] F. Hastarita Rachman, R. Sarno, and C. Fatichah, "Music Emotion Classification based on Lyrics-Audio using Corpus based Emotion," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 3, pp. 1720–1730, 2018.

[9] F. H. Rachman, R. Sarno, and C. Fatichah, "CBE: Corpus-based of emotion for emotion detection in text document," in *Proceedings - 2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2016*, 2017, pp. 331–335.

[10] B. S. Rintyarna, R. Sarno, and C. Fatichah, "Enhancing the performance of sentiment analysis task on product reviews by handling both local and global context," *International Journal of Information and Decision Science*, vol. 11, 2018.

[11] E. W. Pamungkas, R. Sarno, and A. Munif, "B-BabelNet: Business-Specific Lexical Database for Improving Semantic Analysis of Business Process Models," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 15, no. 1, p. 407, 2017.

[12] R. Navigli and S. P. Ponzetto, "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network," *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.

[13] B. S. Rintyarna and R. Sarno, "Adapted weighted graph for Word Sense Disambiguation," in *2016 4th International Conference on Information and Communication Technology, ICoICT 2016*, 2016.

[14] S. Banerjee and T. Pedersen, "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet," *International Conference on Intelligent Text Processing and Computational Linguistics*, vol. 2276, pp. 136–145, 2002.

Alignment Model of Quality Assurance System of Higher Education And Performance Measurement Based on Framework CobiT 5

Budi Widjajanto
Department of Information System
Dian Nuswantoro University
Semarang, Indonesia
bdpojok @gmail.com

Dewi Agustini Santoso
Department of Information System
Dian Nuswantoro University
Semarang, Indonesia
dewi@dsn.dinus.ac.id

Nova Rijati
Department of Informatics
Dian Nuswantoro University
Semarang, Indonesia
nova.rijati@dsn.dinus.ac.id

Abstract — *This study proposes an Alignment Model of IT Governance and Performance Measurement based on Framework CobiT 5, which is conducted by identifying Business Goals, IT Goals, IT Process and Framework Alignment. The research results indicate that the education quality standard of Dian Nuswantoro University which fulfill Regulation of Ministry of Research, Technology and Higher Education No. 44 of 2015 is 42% and the remaining 58% has not fulfilled it. Model of IT Governance as a University Quality Assurance System with ISO/IEC 38500: 2008 and Process Reference Model CobiT 5 frameworks could be performed by aligning Dian Nuswantoro University's goals to the Process Reference Model CobiT 5 and adjusted with the IT governance principles based on ISO/IEC 38500 framework. However, the capability level of IT processes is mostly at level 1, showing that the IT processes have indeed been implemented and some of them have reached level 2, indicating that the processes have been implemented and managed. Meanwhile, it is expected that the capability level could mostly be at level 4, an indication that the IT processes should be implemented, managed and standardized and a performance measurement should be performed.*

Keywords — Model of IT Governance, Quality Assurance System, ISO/IEC 38500: 2008 Framework, Process Reference Model COBIT 5

I. INTRODUCTION

The implementation of information technology governance in an organization, including universities, requires a help from some tools in the form of a comprehensive information technology governance framework, aiming at providing guidelines for leaders in managing their information technology well within the organization they are leading [1][2][3][4][5]. A university's IT governance model could be built using ISO 38500 framework with its governance components being determined based on Cobit 5 enabler. The components used include goals, three pillars (tri dharma), conformance and performance of universities. However, the university's IT governance model proposed frequently disregard the aspects which has an effect on how the activities in implementing the three-pillar components are suitable, i.e. related to the quality assurance system, which aims at maintaining and improving Higher Education quality sustainably in implementing the university's three-pillar components[6][7].

Several frameworks of information technology governance include Cobit, ITIL, ISO, etc. Cobit is the most commonly used framework for information technology governance[8][9] which comprehensively helps the organization achieve the goals through information technology governance and management[10]. Cobit also provides guidelines on how information technology governance should be built by considering the company governance and its scope of information technology governance. The conceptual model of information technology governance is built based on the main principles whose existences are a must in a university governance process. This way, it will serve as a guideline for higher education to develop a blueprint of information technology which does not just support the information technology aspect of academic and non-academic activities, but also takes into account the entire scope of university governance. Meanwhile, ISO/IEC 38500: 2008 is a standard for information technology governance which provides principles for senior management to evaluate, direct and monitor the use of information technology [11].

In order to implement Regulation of the Minister of Research, Technology and Higher Education No. 62 of 2016 concerning Higher Education Quality Assurance System and to improve university's competitiveness to achieve its competitive excellence in the face of ASEAN Economic Community, then, this research proposes an information technology governance model as the basis of University Quality Assurance System by applying the principles of ISO/IEC 38500: 2008 on which the quality assurance system is based and Cobit 5 framework as a guideline in implementing information technology processes which have been aligned with the university's quality assurance processes. It is expected that the university's information technology governance model proposed here would be capable of maintaining and improving its quality sustainably in performing the university's three-pillar components as well as increasing its competitiveness in the face of ASEAN Economic Community.

II. Review of Literature

i. IT Governance Model

IT governance is "The responsibility of the board of directors and executive management, which consists of leadership, organizational structure and processes that ensure that the company's IT supports and broadens the company's strategies and objectives"[12]. Additionally, IT governance has been of great importance in a business process in an organization in order to ensure that IT has been managed well which drives the organization to control IT related to the IT values obtained and preventive control over the risks occurring from the implementation of such IT [9]. IT governance is also an integral part of an enterprise governance's success through the improvement in efficiency and effectiveness in the company's relevant processes, and IT governance also provides a structure which connects IT processes, IT resources and information to the company's strategy and goals. Moreover, IT governance combines best practices from the IT planning and organization, development, and implementation, delivery, and support as well as monitors the IT performance to ensure if the company's relevant information and technology support the company's business goals [12].

ii. Higher Education Quality Assurance System

Higher Education Quality Assurance System is a set of activities to improve Higher Education quality in a well-planned and sustainable way [13] through three subsystems which are synergizing one another in their implementations, i.e. (1) Indonesian Higher Education Database, (2) Internal Quality Assurance and (3) External Quality Assurance System to fulfill or exceed the National Education Standards in the effort of continuously assuring quality, as can be seen below [14]:

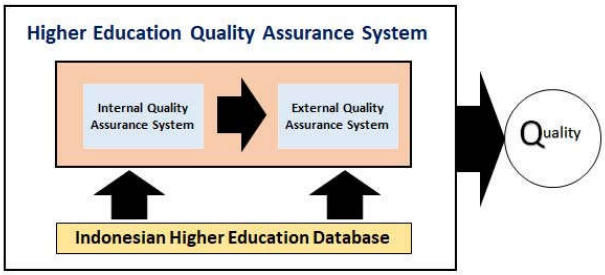


Figure 1. Higher Education Quality Assurance System

University quality here is related to the level of conformance between higher education administration and the Higher Education Standards consisting of National Higher Education Standards stipulated by the Agency of National Education Standards and Higher Education Standards established by the University, under the Regulation of the Minister of Research, Technology and Higher Education No. 62 of 2016 article 3, paragraph 1 stating that Higher Education Quality Assurance System comprises of Internal Quality Assurance System and External Quality Assurance System.

iii. CobiT 5 Framework

CobiT is a framework of IT governance developed by Information Systems Audit and Control Association (ISACA) where the initial development of CobiT framework is to be a framework to perform IT audit tasks known as "Control Objectives for IT Processes". Later on, CobiT framework keeps on developing and has been used as a more expanded IT governance and management [10]. CobiT 5 provides a comprehensive framework in helping companies reach their goals in their IT governance and management. It helps the company in creating the optimal value of IT while maintaining the balance between the benefits it realizes by optimizing the risk level and resources use. CobiT 5 framework enables IT to be arranged and managed holistically for the whole company, taking in the full end-to-end business and functional areas of responsibility, considering the IT-related interests of internal and external stakeholders. Below are the principles of CobiT 5 framework [15]:

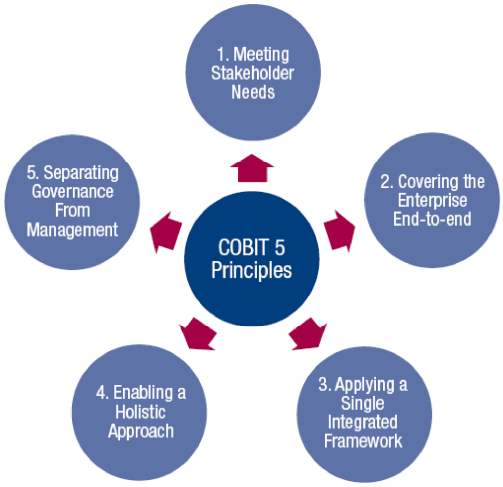


Figure 2. CobiT 5 Principle

In the CobiT 5 framework, the process reference model clearly distinguishes governance from management where they are classified into domains as follows [16]:

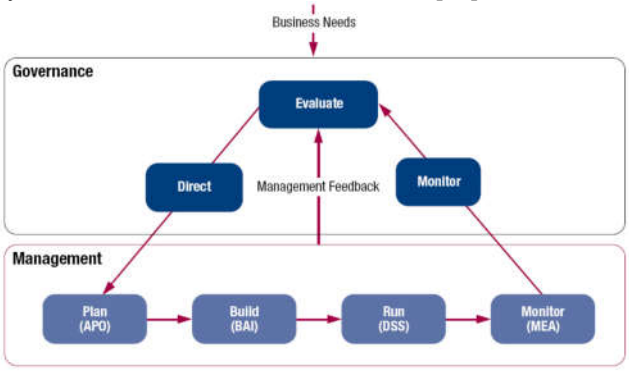


Figure 3. The separation between Governance and Management

From this, CobiT 5 is then divided into 5 domains, where 4 domains belong to management process, namely *Align, Plan and Organise* (APO), *Build, Acquire and Implement* (BAI),

Deliver, Service and Support (DSS) and *Monitor, Evaluate and Assess* (MEA) and 1 domain belongs to governance, namely *Evaluate, Direct and Monitor* (EDM). In CobiT 5 there are 37 processes in total which are classified into 5 domains with two governance and management areas.

III. Research Method

The research is conducted in Dian Nuswantoro University (UDINUS), Semarang. The sample is taken using purposive sampling technique based on RACI Chart Cobit 5. The data collected in this research are primary and secondary data obtained from various sources. They are collected through a number of steps namely, primary data are obtained through: interview, observation and survey and the secondary ones are obtained through documentation study, internet access as well as other relevant studies.
Figure 5 is the design of this research which will be conducted:

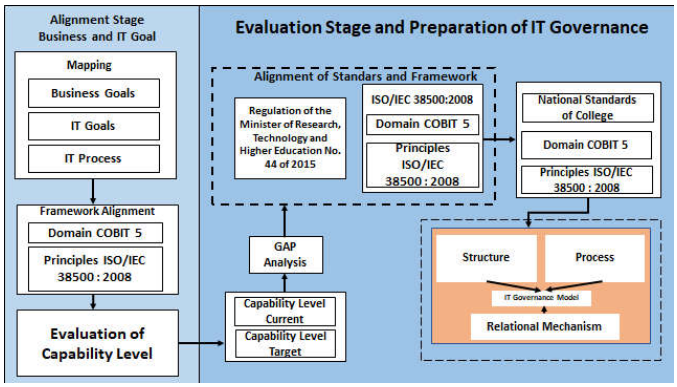


Figure 4. Research Method

IV. Research Results

A. Business Goal Identification

The process of identifying business goals is performed by first defining the organization's goals and targets. After the goals and targets have already been formulated, a mapping process of those goals and targets using the business goals owned by CobiT 5 would then be performed would then be performed [16]. Furthermore, a mapping is made based on the business goals and IT goals in matching the IT processes which support the business goals, by first mapping into the applicable business goals.

No	CobiT 5 Business Goals	BSC Dimension
3	Managing business risks (securing assets)	*Financial*
5	Financial Transparency	
6	Customer-oriented service culture	*Customer*
7	Business service availability and sustainability	
8	Fast response to changing the business environment	
11	Business process functionality optimization	*Internal*
13	Business change management program	
14	Operating and employee productivity	
15	Fulfillment of internal policy	
16	Competent and motivated employee	*Learning and Growth*
17	Product innovation and service culture	

B. IT Goal Identification

IT goal identification is done after the business goal identification process has already been defined. This process is to map business goals with IT goals, where CobiT 5 has mapped between business goals and IT goals, which aims at finding out what IT goals are going to support the organization's business goals later [16]

No	*Business Goals*	*IT Goals (Primary)*
3	Managing business risk (securing assets)	– ITG04. Managing IT-related business risk – ITG10. Information security, application, and infrastructure processing – ITG16. Competent and motivated IT personnel and business
5	Financial transparency	– ITG06. Transparency of IT costs, benefits and risks
6	Customer-oriented service culture	– ITG01. IT and business strategy alignment – ITG07. IT service delivery in line with business needs
7	Business service availability and sustainability	– ITG04. Managing IT-related business risks – ITG10. Information security, application, and infrastructure processing – ITG14.
8	Fast response to changing the business environment	– ITG01. IT and business strategy alignment – ITG07. IT service delivery in line with business needs – ITG09. IT skills – ITG17. Knowledge, expertise, and initiative for business innovation
11	Business process functionality optimization	– ITG01. IT and business strategy alignment – ITG07. IT service delivery in line with business needs – ITG08. Use of adequate application, information, and technology solution – ITG09. IT skills – ITG12. Empowerment and

No	*Business Goals*	*IT Goals (Primary)*
		support of business process by integrating application and technology into business processes
13	Business change management program	– ITG01. IT and business strategy alignment – ITG03. Executive management's commitment to making IT-related decisions – ITG13. Program delivery gives benefit in a timely manner, as budgeted and meets the needs and quality standards
14	Operating and employee productivity	– ITG08. Use of adequate application, information, and technology solution – ITG16. Competent and motivated IT personnel and business
15	Fulfillment of internal policy	– ITG02. Conformance of IT and support to fulfill the business with – ITG10. Information security, application, and infrastructure processing – ITG15. Conformance of IT with internal policy
16	Employees who are competent and motivated	– ITG16. IT and business personnel are competent and motivated
17	Product innovation and service culture	– ITG09. IT skills – ITG17. Knowledge, skills and initiatives for business innovation

C. IT Process Identification

The process is carried out further in aligning business strategies and IT strategy is to conduct a mapping of the IT goals which have previously been identified with the IT process which has some relationship by first focusing the IT processes of Primary predicate in determining the IT processes since the ones with primary relationships are the IT primary processes in supporting the achievement of desired IT goals [16] CobiT 5 framework has 37 IT processes which are classified into 5 domains.

No	*IT Goals (Primary)*	*IT Process (Primary)*
1	IT and business strategy alignment	EDM01, EDM02, APO01, APO02, APO03, APO05, APO07, APO08, BAI01, BAI02
2	IT suitability and support for business compliance with external laws and regulations	APO01, APO12, APO13, BAI10, DSS05, MEA02, MEA03
3	Executive management's commitment to making IT-related decisions	EDM01, EDM05
4	Managing IT-related business risks	EDM03, APO10, APO12, APO13, BAI01, BAI06, DSS01, DSS02, DSS03, DSS04, DSS05, DSS06, MEA01, MEA02, MEA03
6	Transparency of IT costs, benefits and risks	EDM02, EDM03, EDM05, APO06, APO12, APO13, BAI09
7	IT service delivery in line with business needs	EDM01, EDM02, EDM05, APO02, APO08, APO09, APO10, APO11, BAI09
8	Use of adequate application, information, and technology solution	APO04, BAI05, BAI07
9	IT skills	EDM04, APO01, APO03, APO04, APO10, BAI08
10	Information security, application, and infrastructure processing	EDM03, APO12, APO13, BAI06, DSS05
12	Empowerment and support of business processes by integrating applications and technology into business processes	APO08, BAI02, BAI07
13	Program delivery gives benefit in a timely manner, as budgeted and meets the needs and quality standards	APO05, APO07, APO11, APO12, BAI01, BAI05
14	Availability of reliable and useful information for decision-makers	APO09, APO13, BAI04, BAI10, DSS03, DSS04
15	Conformance of TI with internal policy	EDM03, APO01, MEA01, MEA02
16	Competent and motivated IT personnel and business	EDM04, APO01, APO07

No	*IT Goals (Primary)*	*IT Process (Primary)*
17	Knowledge, expertise, and initiative for business innovation	EDM02, APO01, APO02, APO04, APO07, APO08, BAI05, BAI08

D. Process Alignment And Measurement Performance

The combination between CobiT 5 and ISO / IEC 38500: 2008 frameworks needs to be done first because ISO / IEC 38500: 2008 provides a high level of standards for IT governance while CobiT provides additional insight into what can help implement and improve the six principles of ISO / IEC 38500: 2008 in providing IT governance as well as effective business and IT alignment [11].

ISO Principle	**Cobi T 5**	**EDM**	**APO**	**BAI**	**DSS**	**MEA**
Responsibility		EDM 05	-	-	-	-
Strategy		EDM 02	-	-	-	-
Acquisition		-	APO 05	-	-	-
Performance		-	APO 02 APO 09	-	-	MEA 01
Conformance		-	APO 02	-	-	MEA 02 MEA 03
Human Behaviour		-	APO 07	BAI 02 BAI 05 BAI 08	-	-

The results of goal and target alignment being made indicate that the Dian Nuswantoro University's business goals have included four balanced scorecard perspectives where mostly dominated by internal perspective, followed by customer perspective. The alignment of IT process results of CobiT 5 with ISO/IEC 38500:2008 principle along with each capability level can be seen below:

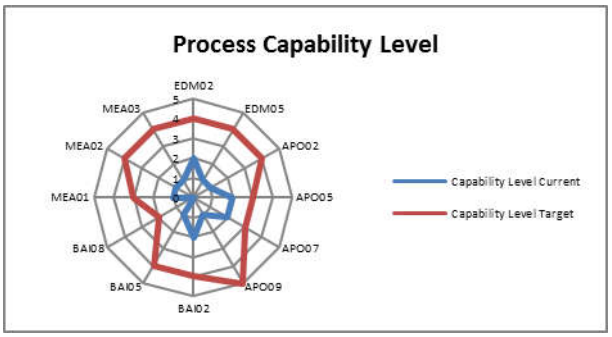


Figure 5. Capability Level of Udinus's IT Processes

It shows that the current Dian Nuswantoro University's capability processes are mostly at level 1, and some have reached level 2, yet there is a process which is still at level 0. Dian Nuswantoro University's capability processes sitting at level 1 are the process of ensuring stakeholder transparency (EDM05), the process of managing strategy (APO02), the process of managing service arrangement (APO09), the process of managing possible organizational change (BAI05), the process of monitoring, evaluating and assessing performance and conformance (MEA01), process monitoring, evaluating and assessing the internal control system (MEA02) and the process of monitoring, evaluating and assessing compliance with external needs (MEA03). Those Dian Nuswantoro University's capability processes sitting at level 2 are the process of ensuring to earn benefits (EDM02), the process of managing a portfolio (APO05), the process of managing human resources (APO07) and the process of managing need definition (BAI02). Lastly, the only Dian Nuswantoro University's capability process remaining at level 0 is the process of managing knowledge (BAI08).

Meanwhile, it is expected that Dian Nuswantoro University's capability processes could mostly sit at level 4 and several capability processes are expected to be at level 3 and only one capability process is expected to be at level 5. Those capability processes expected to be at level 4 are the process of managing knowledge (BAI08), the process of ensuring transparency to stakeholders (EDM05), the process of managing strategy (APO02), the process of managing possible organizational change (BAI05), the process of monitoring, evaluating and assessing internal control system (MEA02), the process of monitoring, evaluating and assessing compliance with external requirements (MEA03), the process of ensuring to earn benefits (EDM02) and the process of managing need definition (BAI02). Meanwhile, Dian Nuswantoro University capability's processes expected to be at level 3 are the process of monitoring, evaluating and assessing performance dan conformance (MEA01), the process of managing a portfolio (APO05) and the process of managing human resources (APO07). Finally, the only Dian Nuswantoro University's process expected to be at level 5 is the process of managing service arrangement (APO09).

From the results of aligning education quality standards owned by Dian Nuswantoro University with Dikti's national standards as stipulated in Permenristekdikti No. 44 of 2 Regulation of the Minister of Research, Technology and

Higher Education No. 44 of 2015, the following chart is obtained:

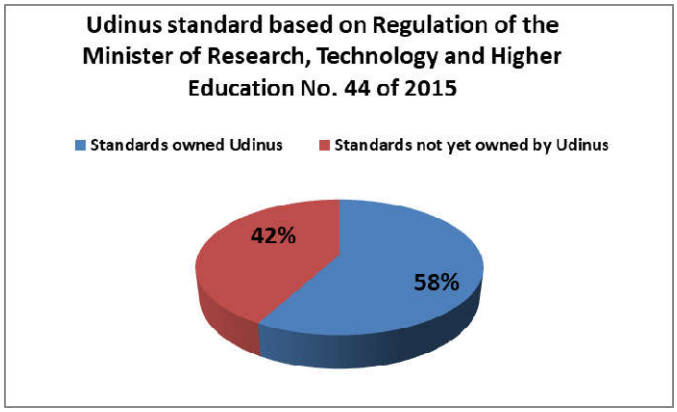


Figure 6. Udinus's Higher Education Standards

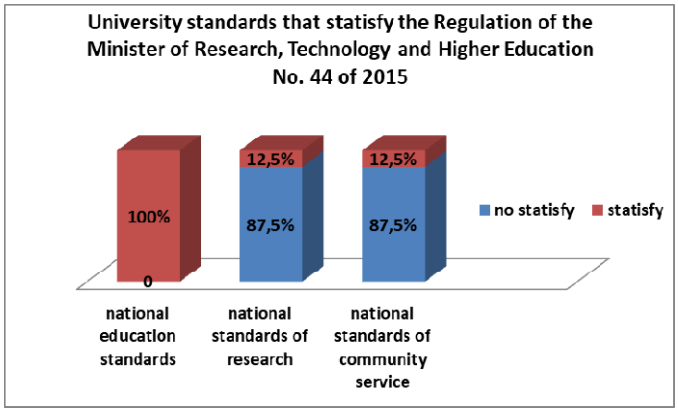


Figure 7. Higher Education Standards

From the results of adjustment between the National Higher Education Standards and higher education quality standards of Dian Nuswantoro University, it is found that the education quality standards owned by Dian Nuswantoro University have not incorporated all of the National Higher Education Standards stipulated by Permenristekdikti which consist of National Education Standards, research and service to community, particularly in the research and community, many of Dian Nuswantoro University standards have not met the standard requirements established by Permenristekdikti. As for the National Education Standards, Dian Nuswantoro University has met the predetermined standard requirements in general. Also, Dian Nuswantoro University has developed higher education quality standards as needed by the internal stakeholders within the University.

The evaluation results of IT Governance Model proposed here show that it is acceptable at 85.22% by Dian Nuswantoro University which includes 88.98% in relation to the proposed IT organizational structure design, 86.67% in relation to IT strategy planning design and 80.00% in relation to the dashboard design in implementing the education quality standards.

V. Conclusion

The research results indicate that 42% of Dian Nuswantoro University's education quality standards has met Regulation of the Minister of Research, Technology and Higher Education No. 44 of 2015, and the remaining 58% has not. The IT Governance Model as a University Quality Assurance System using ISO/IEC 38500: 2008 framework and Process Reference Model CobiT 5 could be implemented by aligning Dian Nuswantoro University's goals with the Process Reference Model CobiT 5 and adapted to the principles of IT governance based on the ISO / IEC 38500 framework. However, most of the IT capability processes are sitting at level 1, indicating that the IT processes have been performed and several processes have even reached level 2, an indication that these processes have been implemented and managed. Meanwhile, it is expected that most capability processes would be at level 4 which shows that the IT processes should be implemented, managed, standardized and measured for its performance.

Reference

[1] G. Mangalaraj, a Singh, and a Taneja, "IT Governance Frameworks and COBIT-A Literature Review," *Twent. Am. Conf. Inf. Syst.*, pp. 1–10, 2014.

[2] H. Tanuwijaya and R. Sarno, "Comparation of CobiT Maturity Model and Structural Equation Model for Measuring the Alignment Between University Academic Regulations and Information Technology Goals," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 10, no. 6, pp. 80–92, 2010.

[3] G. S. Kearns and R. Sabherwal, "Strategic Alignment Between Business and Information Technology: A Knowledge-Based View of Behaviors, Outcome, and Consequences," *J. Manag. Inf. Syst.*, vol. 23, no. 3, pp. 129–162, 2007.

[4] B. H. Reich, I. Benbasat, M. Information, and S. Division, "Factors That Influence the Social Dimension of Alignment Between Business and Information Factors That Influence the Social Dimension of Alignment Between Business and Information Technology Objectives," vol. 24, no. 604, pp. 1–59, 1999.

[5] G. Premkumar and W. R. King, "An empirical assessment of information systems planning and the role of information systems in organizations," *J. Manag. Inf. Syst.*, vol. 9, no. 2, pp. 99–125, 1992.

[6] H. Nugroho, "Conceptual Model of It Governance for Higher Education Based on Cobit 5 Framework.," *J. Theor. Appl.*, no. February 2014, 2014.

[7] H. Nugroho and K. Surendro, "Proposed model of vocational university governance and measurement model by utilizing the ISO 38500 framework and COBIT 5 enabler," *Proc. - Int. Conf. ICT Smart Soc. 2013 "Think Ecosyst. Act Converg. ICISS 2013*, pp. 86–90, 2013.

[8] N. Rijati, D. A. Santoso, and B. Widjajanto, "Integrated strategy of quality insurance system with information

technology process in universities," *Proc. - 2017 Int. Semin. Appl. Technol. Inf. Commun. Empower. Technol. a Better Hum. Life, iSemantic 2017*, vol. 2018–Janua, pp. 31–37, 2018.

[9] G. L. Lunardi, J. L. Becker, A. C. G. Maçada, and P. C. Dolci, "The impact of adopting IT governance on financial performance: An empirical analysis among Brazilian firms," *Int. J. Account. Inf. Syst.*, vol. 15, no. 1, pp. 66–81, 2014.

[10] S. De Haes, W. Van Grembergen, and R. S. Debreceny, "COBIT 5 and Enterprise Governance of Information Technology: Building Blocks and Research Opportunities," *J. Inf. Syst.*, vol. 27, no. 1, pp. 307–324, 2013.

[11] A. Chaudhuri, "Enabling Effective IT Governance: Leveraging ISO/IEC 38500:2008 and COBIT to Achieve Business–IT Alignment," *Edpacs*, vol. 44, no. 2, pp. 1–18, 2011.

[12] I. G. Institute, *Framework Control Objectives Management Guidelines Maturity Models*. 2007.

[13] M. of R. T. and H. E. of the R. of Indonesia, "Regulation of the Minister of Research, Technology and Higher Education No. 44 of 2015," 2015. .

[14] M. of R. T. and H. E. of the R. of Indonesia, "Regulation of the Minister of Research, Technology and Higher Education No. 62 of 2016," 2016.

[15] I. G. Institute, *A Business Framework for the Governance and Management of Enterprise IT*, no. September. 2011.

[16] S. De Haes and W. Van Grembergen, "Enterprise governance of IT and the evolutions in COBIT: An academic perspective," *Lect. Notes Bus. Inf. Process.*, vol. 127 LNBIP, no. April, pp. 88–103, 2012.

Classification of Governer's Public Report from SMS LaporGub Using Naive Bayes Classifier Method

Ardytha Luthfiarta
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
ardytha.luthfiarta@dsn.dinus.ac.id

Rizky Herliana
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
rizkyherliana@gmail.com

Rismiyati
Department of Computer Science
Diponegoro University
Semarang, Indonesia
rismi@if.undip.ac.id

Junta Zeniarja
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
junta@dsn.dinus.ac.id

Heru Agus Santoso
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
heru.agus.santoso@dsn.dinus.ac.id

Desi Purwanti Kusumaningrum
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
desi.purwanti@dsn.dinus.ac.id

Adhitya Nugraha
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
adhitya@dsn.dinus.ac.id

Erwin Yudi Hidayat
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
erwin@dsn.dinus.ac.id

***Abstract*—Freedom to express people's feeling like complaint is one effort from a governor to improve a public service in Central Java Province. By knowing all complaints from people and sometimes they give some suggestions, it able to make better performance government in various aspects such as infrastructure, criminal, healthy , etc. To execute this effort, government establish a "SMS LaporGub" service which able to become an interaction and a communication media that hold all complaints that occur in public. It is already execute well and increase some performance in many aspects. The existence of a wide variety of complaints and high number of reports makes manual analysis becomes ineffective. It needed an automatic classification to classify and analyse the report based on its types. So administrator easy to handle those reports, prevent delay handling for urgent reports to be discussed. This study using classification method especially Naïve Bayes Classifier method**

Keywords — SMS LaporGub, Public's Report, Classification, Naïve Bayes Classifier, Text Mining

I. INTRODUCTION

Governor of Central Java Province showing his concern to realize his vision "Menuju Jawa Tengah Sejahtera dan Berdikari". Several programs have been performed and various innovations have been realized. An example is established Short Message Service (SMS) LaporGub. Despite government already having public services complaints via sites (Laporgub.jatengprov.go.id) and Twitter , the governor is not satisfied and launched a new public service report via SMS LaporGub [1].

SMS LaporGub is a media communication and interaction between government and society [2]. Not only that, it also accommodate a variety of reports such as complaints, suggestions, and gratitude. As complaints and suggestions reports discuss about infrastructure, public services, social aid, energy, criminal, education, health, law, until scarcity of clean water and fertilizer, etc[2]. SMS laporGub can be an evaluation point for measure the performance of every institution especially in public sectors that require special handling.

SMS LaporGub proven has good results with overall amount of all reports from public is 7492 reports started from April 2015 until the second week of September 2016[3]. Incoming reports are dominated by a variety problems such as public services, infrastructure, security, social assistance, and education [3]. Then followed by human labor, energy, government bureaucracy, criminal, economic, and licensing. It based on the information and data that has been given by administrator.

There was so many reports, process for distributing and handling reports were difficult. High number of people's reports from SMS LaporGub is not able to analyse all problems manually. Automatic classification system can help administrator to handle those reports, prevent delay handling for urgent reports to be discussed. Considering the importance classification system for SMS LaporGub, it needed technique to solve it. Text Mining is one the most technique to solve automatic classification. The fundamental step in text mining involves converting unstructured text into a more explicitly structured intermediate format [4].

The purpose of this study is to analyse and identify type of the problems that occur in society. This study aims to establish an automatically classification reports of SMS LaporGub using Naïve Bayes Classifier method.

Amir Hamzah [5] prove that naive bayes classifier has high enough performance for classification news and academic documents. Problem idea of this research is

the difficulty of managing information for news and academic abstract which become one of an important type of document that always increases. For classification news documents have 91% accuracy and for academic documents has 82% accuracy.

Moromi Gogoi and Shikhar Kumar Sarma [6] has been done with their research using naive bayes classifier. The Research was conducted by classifying documents into four classes i.e. sports, politics, law and science. To build and evaluate this classification model, a total of 200 documents is split into two datasets items, namely the training set and testing set, in the which 60% of the documents is used as training set whereas the remaining 40% is used as the testing set. The result of this model has given a standard Precision of Recall of 94.41% and 94.68%.

On the other hand, there are some related works of classification using Naïve Bayes Classifier method. Research by Mahesh Kini M, Saroja Devi H, Prashant G Desai, Niranjan Chiplunkar [7]. Number of documents increase rapidly making a complicated job. The problems that always exist which large document is a hassle when administered in every sector. To solve the problem, authors used naive bayes classifier to manage various types of documents effectively. Result of this research is 96.9% accuracy performance, 0.969 precision,0.969 recall, 0.969F-measure. This result by using preprocessing and feature selection.

Amir Hamzah [6] prove that naive bayes classifier has high enough performance for classification news and academic documents. Problem idea of this research is the difficulty of managing information for news and academic abstract which become one of an important type of document that always increases. For classification, news documents have 91% accuracy and for academic documents has 82% accuracy. Moromi Gogoi and Shikhar Kumar Sarma [6] has been done with their research using naive bayes classifier. The Research conducted by classifying documents into four classes i.e. sports, politics, law and science. To build and evaluate this classification model, a total of 200 documents is split into two datasets items, namely the training set and testing set, in the which 60% of the documents is used as training set whereas the remaining 40% is used as the testing set. The result of this model has given a standard Precision of Recall of 94.41% and 94.68%.

II. TEXT CLASSIFICATION PROCESS

A. Classification

Classification is performed to classify class labels of a data instance into a prespecified set of categories [8]. Classification process consists of two-step process, training process and testing process [9].

Based on Fig. 1, x is data, and y is the label. (x,y) can be used to develop model during training process. It can be used to predict class label (y) from data test (x,?). Learning algorithm is needed for this training process to implement the model on data test. It will give predictive class.

Classification must be measured to test the accuracy of the model. A set of example is given to measure how correctly classified frequently as class x or misclassified as some other class. This evaluation model classifier is given in a confusion matrix.

From that confusion matrix can be measured accuracy analyze a performance of classification algorithm. Accuracy value represent how much the entire document is classified correctly. This evaluation model classifier is given in a confusion matrix as shown in Table 1[10].

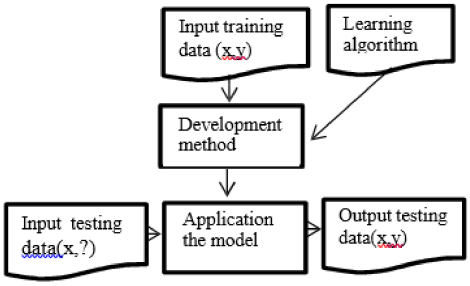


Fig. 1. Classification Framework

TABLE I. CONFUSION MATRIX TABLE

Classication		Predicted Class	
		Class= Yes	Class= No
Observed Class	Class= Yes	a	c
	Class= No	b	d

a = frequent true positive classification where correctly labeled by the classifier.

b = frequent false positive classification in negative tuples where incorrectly labeled as positive.

c = frequent false negative classification is positive tuples that were mislabeled as negative.

d = frequent true negative classification is negative tuples were correctly labeled by the classifier.

Formula of accuracy defined as shown in (1).

$$accuracy = \frac{(a+d)}{(a+b+c+d)} \quad (1)$$

Based on accuracy value, it will be classified into some categories [10]:

0.90 - 1.00	Excellent classification
0.80 - 0.90	Good classification
0.70 - 0.80	Fair classification
0.60 - 0.70	Poor Classification
0.50 - 0.60	Failure

B. Text Mining

Text Mining is a process of extracting useful information from document collection of textual data [11].

The application of text mining can use in many aspects such a filtering spam documents, measuring customer preferences, summarizing documents, grouping research topics, etc.
To obtain the result of text mining, the common steps should be done are text preprocessing and representation vector using Term Frequency Inverse Document Frequency (TF-IDF) weighting [8][12].
Text preprocessing is important part of text mining which responsible for transforming unstructured data into structured data [11]. Preprocessing consists of several stages are [13] case folding (to converts all letters to lowercase and remove another characteristics), tokenizing or parsing (the stage for breakdown all the words to sequence), filtering (process to select the important words, what words are used to represent document), and stemming (a technique for reduction of words into their root word). Learning algorithms cannot directly process the document text in its original form. Therefore, after the pre-processing stage, the document is converted into a more manageable.

C. *Naïve Bayes Classifier*

Naive Bayes Classifier is an algorithm to find the highest value of the probability to classify the testing data to the most appropriate category [10]. The approach of naïve bayes classifier will give the result in highest category label probability (CMAP) with input attributes (w_1,w_2,…,w_i) where w_1is the first word, w_2 is the second word and so on. While C is the set of categories of documents.

III. RESEARCH METHODOLOGY

Fig.2. Shows overview of the research flow in this study. The first step performed in this study is Data acquisition, followed by data pre processing. After data has been cleaned, data analysis is performed to extract features. Finally, data is classified by using Naïve bayes Classifier.

A. *Data Acqusition*

Data used in this study is taken from BIRO HUMAS SETDA in Central Java Province, Indonesia. In total, there are 1100 records, each of which consist of eight attributes. The eight attributes in the data are complain number, content, respond, PIC, location, date, time and category. From the eight attributes, only complain content and category are used in this analysis. In total, there are 11 categories. Those categories are criminal, economic, educational, energy, government bureaucracy, human labor, infrastructure, licensing, public services, security, and social assistance.
Examples of data used for this study is as shown in Table II.

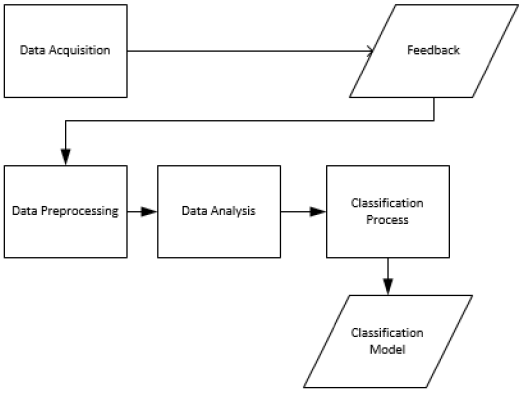


Fig. 2. Overview of the research flow of this study

B. *Data Preprocessing*

Fig. 3. Shows data preprocessing steps used in this study. The first step if data preprocessing is case folding. This process transforms letters into lowercase and removes all punctuation symbol. The next process is tokenizing, which will breakdown all sentence into a token or a single word sequence. Those words will go through filtering process to eliminate unrelated and not meaningful words. The last step is stemming which will change the term into basic form by omitting the word affixes. The stemming is done by using Nazief Adriani method.

TABLE II. DATA EXAMPLE

Content	Category
Selamat siang bpk/ibu.. perkenalkan nma sya: ibu janu prasetiyani,almat gunungwulan kedungjati. mhon penjelasan untuk syarat2 agar ank sya mendapat beasiswa.. Sbelumnya sya ucapkan bnyk terimaksih,atas per hatiannya..	Social assistance
Meminta pd pak gub.. Untk di bantu dalam hal pembibitan penggemukan sapi.. Dan km siap untk menjadi pelopor desa penghasil daging.. Pak..gub.Yg km hormati pak gub.. Jateng, dlm hal program pemberdayaan masyrakat, dan juga meningkatkan ekonomi lokal.. Km wrga kalangbancar kec:geyer.kab:grobogan.	Economy
bagaimana pak masalah penambangan pasir di desa tumiyang kecamatan kebasen kab banyumas apakah d tindak lanjuti…saya mohon segera untuk di tindak lanjuti masalahnya sangat merusak lingkungan dan jalan yang dilewati truk setap hari jadi hancur.terima kasih	Energy

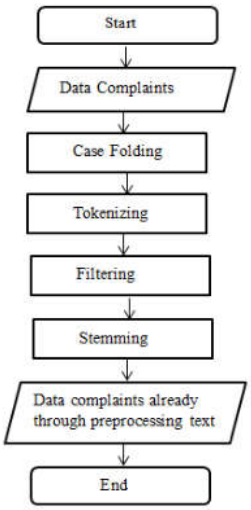


Fig. 3. Data Preprocessing Steps

C. Data analysis and Classification

Clean data resulted from pre-processing process is further analyzed to obtain features. The features used in this process is TF-IDF. Fig. 4. Shows the analysis flow of the data analysis and classification process. As shown in fig. 4, the data for the experiment is split into two categories, training data and testing data. Both training data and testing data will undergo preprocessing and analysis to get TF-IDF.

The training data will undergo training steps as follows:
1. Training data after success through preprocessing text and vector representation using TF-IDF weighted.
2. Calculate the probability of document in every category using equation (2)

$$p(c_j) = \frac{|class\ j|}{|doc\ j|} \quad (2)$$

3. Calculate the probability of every category with specified word by using (3).

$$p(w_i|c_j) = \frac{count\ (w_i|c_j) + 1}{count(c) + |v|} \quad (3)$$

The training steps will result in classification model, in this case, statistics probability, which can be used to classify test data. The test steps for test data are as follows:

1. Doing preprocessing text and vector representation using TF-IDF weighted
2. Calculate $p(w_i|c_j) * p(c_j)$ to every category
3. Find the maximum of $CMAP\ \prod_i P(w_i|c_j) * P(c_j)$

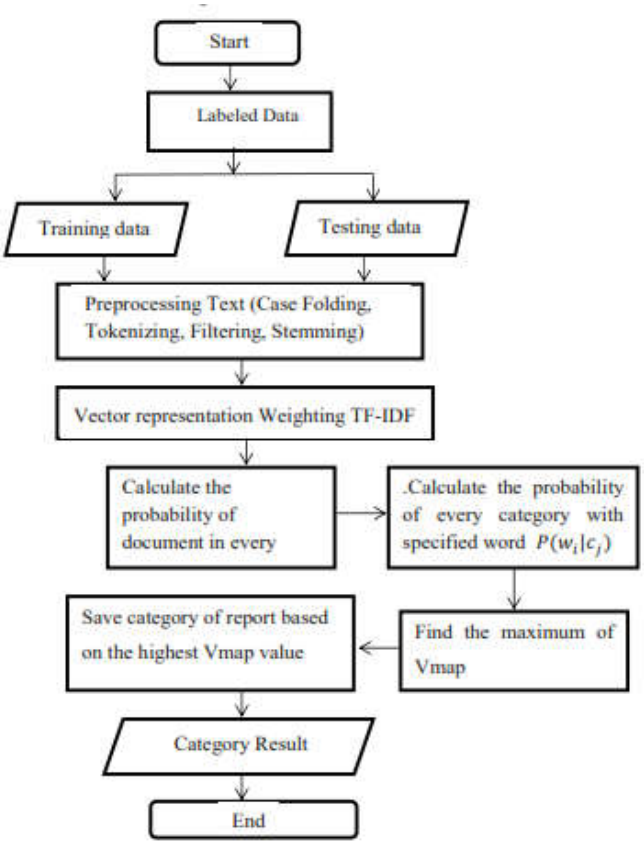


Fig. 4. Data analysis and Classification

IV. Result and Discussion

In this section, the result of experiment is presented and discussed. The performance of the model in this study is measured by 10-fold cross validation. The experiment were performed by using different number of training samples, 550, 880, and 1100. The result of the experiment is shown in in Table III. It is shown from the result that the best accuracy obtained with number of sample of 880. However, the difference in accuracy is not significant since naïve Bayes Algorithm is statistical method and all the number of samples are far above the minimum number of statistic normal distribution. The detailed confusion matrix for the experiment with 880 samples are shown in Table IV. The label conversion for the confusion matrix are shown in Table V. Experiment Result

TABLE III. Experiment Result

No	Amount of Data	Total of True Predictive Data	Accuracy
1	550	482	0.876363
2	**880**	**772**	**0.877272**
3	1100	964	0.876363

TABLE IV. CONFUSION MATRIX OF 880 DATA

	a	b	c	d	e	F	g	h	I	j	k
a	67	0	7	0	0	0	1	0	1	2	2
b	1	64	0	5	0	0	1	1	4	3	1
c	1	0	74	0	1	0	0	0	1	2	2
d	2	1	1	72	0	1	1	1	1	0	0
e	0	0	0	1	74	2	1	0	0	2	0
f	0	0	0	2	0	73	2	0	0	2	1
g	0	0	0	1	1	0	72	2	0	3	1
h	1	1	0	2	1	2	2	66	3	1	1
i	1	2	0	1	1	0	0	0	75	0	0
j	0	0	1	0	2	0	7	3	0	67	0
k	3	1	3	0	1	0	0	2	1	1	68

TABLE V. CODE OF CATEGORY

Variable	Category
a	social assistance
b	government bureaucracy
c	Economy
d	Criminal
e	Infrastructure
f	Security
g	Energy
h	public services
i	Education
j	Licensing
k	human labor

As shown Table above we can conclude that the best experimental result were obtained on test with a total of 880 data. This result is more affected by the data complexity that does not increase even though the data has been added to 1100 data.

V. CONCLUSION

In this paper, LaporGub data classification has been performed. Input data are complain data from LaporGub system. In total, 1100 data from 11 category were used for this study. Preprocessing and data analysis were performed on the data to clean and to produce TF-IDF vector. Naïve Bayes Classifier is used to classify the data. Based on the experiment result, Naïve Bayes Classifier can be applied to classify the governor's report of SMS LaporGub which can help administrator to classify automatically with several process inside.. In addition, the implementation of Naïve Bayes Classifier (NBC) for classifying goversnor's report of SMS LaporGub shown good result of 87.7% accuracy. Further research can be done by adding more data to improve accuracy of the classifier. In addition, the data can be classified by using other classifier such as SVM or neural network.

VI. REFERENCES

[1] metro jateng. (24, April 2015). Metro Jateng . Retrieved November 12 , 2016, from Ingin Wadul ke Gubernur Ganjar? SMS ke Nomor ini: http://metrojateng.com/2015/04/24/ingin-wadul-ke-gubernur-ganjar-sms-ke-nomor-ini/

[2] Humasjtg3. (24, April 2015). *Jatengprov.go.id*. Retrieved November 12, 2016, from HP Jadul Pun Bisa Untuk Lapor Gubernur: http://jatengprov.go.id/id/berita-utama/hp-jadul-pun-bisa-untuk-lapor-gubernur

[3] Humasjtg3. (15, September 2016). *Jatengprov.go.id*. Retrieved November 13, 2016, from Masyarakat Banyak Mengadu Soal Layanan Publik: http://www.jatengprov.go.id/id/berita- utama/masyarakat-banyak-mengadu-soal-layanan-publik

[4] Berry, M. W., & Kogan, J. (2010). *Text mining :applications and theory*. Chichester: Wiley.

[5] Hamzah, A. (2012). Klasifikasi Teks dengan Naive Bayes Classifier (NBC) Untuk Pengelompokan Teks Berita dan Abstract Akademis. *Prosiding Seminar Nasional Aplikasi Sains & Teknologi(SNAST) Periode III ISSN:1979-911X*.

[6] Mangal, S. B., & Goyal, D. V. (2014). Text News Classification System using Naïve Bayes Classifier. *Vidya Publications* (pp. 209-213). An International Journal of Engineering Sciences.

[7] M, M. K., H, S. D., Desai, P. G., & Chiplunkar, N. (2015).

Text Mining Approach to Classify Technical. International Journal of Advanced Research in Computer and Communication Engineering Vol 4 .

[8] Miner, G., Delen, D., Elder, J., Fast, A., Hill, T., & Nisbet, R. (2012). Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications. Oxford: Elsevier.

[9] Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). Fundamentals of Predictive Text Mining. London: Springer.

[10] Asian, J., & Tahaghoghi, H. E. (2005). Stemming Indonesian. *proceedings of the 28 australasian computer*

science conference (ACSC2005).

[11] J. Feldman, R & Sanger, "The Text Mining Handbook :

Advanced Approaches in Analyzing Unstructured Data.,"*Cambridge Univ. Press New York*, 2007.

[12] Weiss, S. M., Indurkhya, N., & Zhang, T. (2010). *Fundamentals of Predictive Text Mining*. London: Springer.

[13] Triawati, C. (2009). *Text Mining*. Retrieved October 27,2016, from Digital Library Telkom Institute of Technology.

A User-centered Design for Redesigning E-Government Website in Public Health Sector

An Approach to Improve the User Experience

Ira Puspitasari
Information System Study Program
Universitas Airlangga
Surabaya, Indonesia
ira-p@fst.unair.ac.id

Dwi Indah Cahyani
Information System Study Program
Universitas Airlangga
Surabaya, Indonesia
dwiindah922@gmail.com

Taufik
Information System Study Program
Universitas Airlangga
Surabaya, Indonesia
taufik-rahman@fst.unair.ac.id

***Abstract*— The rising awareness of information transparency and the increasing citizen participation in "healthy living for healthy societies" campaign have promoted the effectuation of e-government service in public health sector. As one of the government agencies, Surabaya Health Department (SHD), Indonesia, is responsible to provide public health services and to administer health-related policies in Surabaya city. The department has been utilizing an official website to deliver its tasks and responsibilities. However, the website was built mostly based on the department's perspective and internal consensus, without a proper requirement gathering from users and stakeholders. This subjectivity has caused some essential unfulfilled, thus nullified the website goals.**

This study aims to redesign the official SBD website to improve the information dissemination quality and the user experience. The website development follows user-centered design (UCD) methodology. The output consists of user requirements, user task analysis, and a redesigned SHD website prototype. The heuristic evaluations between the existing and the new redesigned websites demonstrate the improvement of user experience. The UCD also enables the user identification and the corresponding requirements more comprehensively. One of the examples is the inclusion of specific requirements for person with color-blind and person with dyslexia in the redesigned SHD website.

Keywords—user-centered design; UCD; usability; public health website; heuristic evaluation; user experience

I. INTRODUCTION

The World Wide Web is arising as one of the major health information sources. Consumers of non-medical professionals, have been extensively utilized health websites to fulfil their health information needs, including access to government policy on health and social care information. The increasing citizen participation in "healthy living for healthy societies" campaign, especially via electronic system, has promoted the effectuation of e-government services in public health sector. Government agencies have developed official websites as one of the core media to deliver the public health services, such as informing verified public health information, promoting healthy life style, preventing the spread of communicable disease, and assuring adequate local public health services.

Delivering public health information electronically requires meticulous and advanced information processing, rigorous evaluation, and strict compliance to standards. To effectively facilitate the presentation of health information, it is essential to design the system based on the user's needs and to incorporate interaction design and usability principles [1]. Previous works also mentioned that usability positively influenced the user satisfaction on mobile-based applications [2], perceived usability affected the continuance usage intention of Internet Banking [3]. In mobile health applications, implementing adequate usability on its development is crucial to ensure the correct exchange of health messages [4]. Delivering accurate health information in the correct presentation is important to avoid misunderstanding of health information as it may lead to harmful consequences [5], e.g., affecting a person's health, or worsening the patient's condition.

Incorporating usability principles and interaction design in a system development is mostly conveyed via its interface design. A good interface enables the users to perform their intended tasks effectively and efficiently without too much distraction. Some of the interface system design methodologies are User-Centered Design (UCD) and Activity-Centered Design (ACD). UCD works based on the needs and the interests of users with a focus of making the system usable and understandable [6], while ACD focuses on the activities that the users will carry out or should be able to perform by using the designed system [7]. ACD is mostly used to design a system where the user groups are heterogenous and the usage goals are varied, but the activities that the users carry out are more common [7]. This study applies UCD because a health-related system must accommodate users' requirements thoroughly to ensure the delivery of relevant and correct health. Therefore, involving user throughout the system development is essential [8]. The UCD design process consists of specifying the context of use, specifying the user and organizational requirements, designing the solution, and evaluating the solution against the requirements [9].

As one of the government agencies, Surabaya Health Department in Indonesia is responsible to provide public health services and to administer health-related policies in Surabaya city. The department has been utilizing an official website to deliver its tasks and responsibilities to Surabaya society. However, the website was built mostly based on the department's perspective and internal consensus, without a proper requirement gathering from users and stakeholders. This subjectivity has caused some essential requirements in a public health system unfulfilled, thus nullified the website goals. Often times, the users were unable to find the public health information they were looking for. Given the current problems in the existing official website, this research aims to redesign the website of Surabaya Health Department to improve the information dissemination quality and the user experience. The proposed solution is developed based on UCD methodology.

II. Research Methodology

This research adopted user-centered design methodology to redesign the official website of Surabaya Health Department. The method consisted the following four phases. The first phase was specifying the context of use by identifying the user and assigning the context of use. To identify the user, we conducted literature study and interviewed the manager of the website from Surabaya Health Department. The purpose of the interview was to gather requirements from the public health service provider's perspective. To assign the context of use, we evaluated the existing websites based on the users' and experts' perspectives using heuristic evaluation. The questionnaire for the heuristic evaluation was developed based on the application of heuristic evaluation in [10], [11]. The questionnaire consisted of ten heuristic indicators, i.e., aesthetic and minimalist design; visibility of system status; match between system and the real world; recognition rather than recall; consistency and standards; flexibility and efficient of use; use control and freedom; error prevention; help users recognize, dialogue, and recovers from errors; and help and documentation. The indicators were measured using 4-point Likert scale, i.e., strongly disagree (1), disagree (2), agree (3), and strongly agree (4). The participants were the users of the existing website, two information system analysts, two public health practitioners, two visual communication designers, and two psychology students.

The second and the third phases were related to the UCD system development. The second phase, specifying the user requirements, identified the user needs and analyzed the user tasks from a public health website. The output of this phase included requirements specification, user tasks analysis, use case diagram, and sequence diagram. The third phase, designing the solution, transformed user requirements into the system blueprint that consisted of the layout design, a storyboard design, and the user interface mock-up. The solution in this study focused on the interaction design via the system interface.

The last phase was evaluating the proposed solution against the requirements using heuristic evaluation method. The evaluators were the same participants who assessed the existing website of Surabaya Health Department in the first phase. The research methodology in this study is depicted in Figure 1.

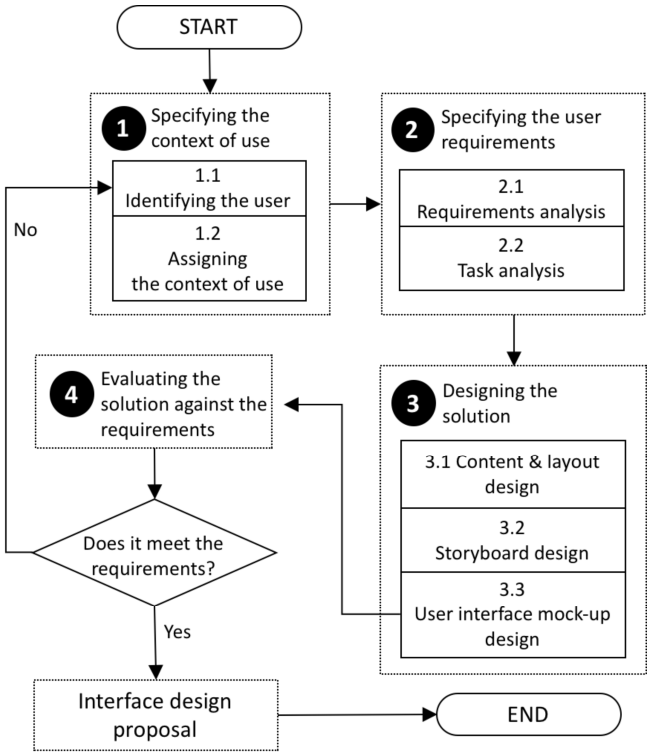


Fig. 1. Research methodology

III. Results: The Website Design Based on UCD

A. User Identification

Based on the interview with the manager of Surabaya Health Department website and literature study in public health, there were three user classes as elaborated in Table I. Each user classes used the website for specific purpose and carried out different tasks.

TABLE I. User Identification

User class	Profession	Organizational Environment	Usage Purpose
Health service provider	1. Hospital staff 2. Community Health Center staff	Hospital, CHC, Specialist Clinics, Drugstore, University	Input health report
	3. Physician / midwife 4. Pharmacist 5. Nutritionist 6. Health analyst 7. Academics		Verify health information, provide public health service, administer online health-related license application
Government stakeholder	1. City government 2. City council	Government office	Monitor the health service provider performance, draw and issue public health policy
Society	1. Citizen 2. Journalist	The city of Surabaya, TV studio, Radio station, News agency	Access public health and healthcare facility information, consult public health matter, download health data.

B. *Assigning the Context of Use (Heuristic Evaluation 1)*

Assigning the context of use involved collecting the user feedback and evaluating the usage of the existing Surabaya Health Department website. Each heuristic indicator consisted specific questions. All answers from all participants for each question was tabulated using the behavior measurement scale [12]. The result of each question was then interpreted based on four interval scales, as shown in the Likert scale. The result of the heuristic evaluation is elaborated in Table II.

TABLE II. THE RESULT OF EXISTING WEBSITE EVALUATION (HEURISTIC EVALUATION 1)

Heuristic Indicator	Question Code	Tabulated Score	Interval Interpretation
Aesthetic and minimalist design	AE1	50%	Disagree
	AE2	55%	Agree
	AE3	46%	Disagree
Visibility of system status	VI1	48%	Disagree
	VI2	53%	Agree
	VI3	58%	Agree
	Vi4	65%	Agree
	VI5	48%	Disagree
Match between system and the real world	MA1	56%	Agree
	MA2	43%	Disagree
	MA3	70%	Agree
Recognition rather than recall	RE1	55%	Agree
	RE2	50%	Disagree
Consistency and standards	CO1	53%	Agree
	CO2	48%	Disagree
	CO3	63%	Agree
Flexibility and efficient of use	FL1	46%	Disagree
	FL2	50%	Disagree
Use control and freedom	US1	51%	Agree
	US2	48%	Disagree
Error prevention	ER1	55%	Agree
	ER2	48%	Disagree
Help users recognize, dialogue, and recovers from errors	HE1	56%	Agree
Help and documentation	HD1	45%	Disagree

C. *Requirement Analysis*

The user requirement analysis incorporated all participants' feedbacks in the heuristic evaluation 1. The result is shown in Table III.

TABLE III. THE PROPOSED REQUIREMENT BASED ON USER'S FEEDBACK FROM THE HEURISTIC EVALUATION 1

No.	Proposed Requirement	Participant
1.	User interface redesign	P1, P3, P4, P7, P8, P9, P10, P11, and P12
2.	Search tools to seek information about health care service	P1, P4, P5, P10, P12, P13, and P15
3.	Direct and real time communication between the user and the staff of Surabaya Health Department	P1, P3, P10, and P11
4.	Inclusion of health data and public health documentation in the website	P2 and P4
5.	Online processing of clinic operating permit	P7, P12, and P15
6.	A user's feedback facility	P4

D. *Task Analysis*

The users' tasks analysis integrated the user identification from the service provider perspective and the requirements from the user's perspective. Table IV shows the task analysis result for each defined user class.

TABLE IV. THE USERS' TASKS

User Class	Task	
Health service provider	1.1	Access Surabaya Health Department website
	1.2	Supply verified and accurate public health information (including health-related news, health-care facilities, and public health policy)
	1.3	Seek public health information
	1.4	Administer online health-related license application
	1.5	Upload health data and public health report / documentation
Govern-ment stake-holder	2.1	Access Surabaya Health Department website
	2.2	Seek public health information
	2.3	Monitor the Surabaya Health Department activities by assessing the posted content related to public health activities in Surabaya city.
Society	3.1	Access Surabaya Health Department website
	3.2	Seek public health information
	3.3	Access public health service
	3.4	Consult public health matter with the staff of Surabaya Health Department
	3.5	Submit feedback related to the website and public health activity
	3.6	Download health data and public health report / documentation

The task analysis also produced use cases, scenario, and the sequence diagram. The use case consisted of three actors and eleven use cases.

E. *Designing the Solution: the Content Design*

The first activity of designing the solution was constructing the content design. Content should be constructed consistently, it defines the layout, the structure, and the outline of all website

content. In this study, the content design focused on the aesthetics and the navigational design.

1) Aesthetics Design

The aesthetic design included the theme colors, layout, and font-schemes. The selected theme colors for the proposed new website were blue (hex code: #0088cc), black (hex code: #383f48), and white (hex code: #FFFFFF). We selected the theme colors to accommodate the color-blind users in the society of Surabaya. About 95% of color-blind population is unable to distinguish green and red colors. Surabaya Health Department also requested the blue color as it is the institutional color.

The website layout is presented in Fig. 2. The main menu, that showed the highest level of user interaction, was located on the top of the page and the side menu (if available) was located on the left part. Since the redesigned website belongs to website with specific topics and interest group, placing the top priority content on the top menu enables greater exposure. This allows the user to find the most important website content faster and easier. The submenu on the left page also facilitated quick eye scanning about the overall structure of a website [13]. The submenu position is more preferable on the left side than on the right side because it is more suitable with the reading direction in Indonesian language and English.

In this first website redesign, the user requirements included specific needs for people with color blind and people with dyslexia. According to International Dyslexia Association, around 10% of people have dyslexia [14]. Thus, we selected the font-schemes with the highest readability to facilitate reading for person with dyslexia, i.e., Helvetica, Arial, and Verdana [14].

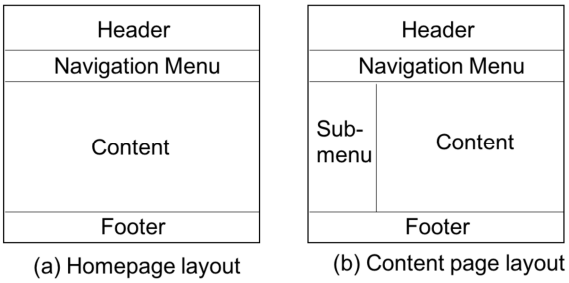


Fig. 2. The layout-design

2) Navigational Design

The website navigation was placed on the top page and constructed based on the hierarchical structure. The top menu consisted of Home, *Tentang Kami* (profile), *Layanan Publik* (public service), *Informasi Publik* (public information), *Download* (health data and report), and *Pojok Kesmas* (*Community Health Center Corner*). Some of the main menu can be decomposed further into several submenu, such as the *Layanan Publik* menu. *Layanan Publik* consisted of E-health, *Pelayanan Puskesmas* (Community Health Center service), *Pelayanan Sarana Kesehatan* (public health facility service), *Perijinan Online* (*online health-related license application*). Fig. 3. depicts the hierarchical structure of the web navigation.

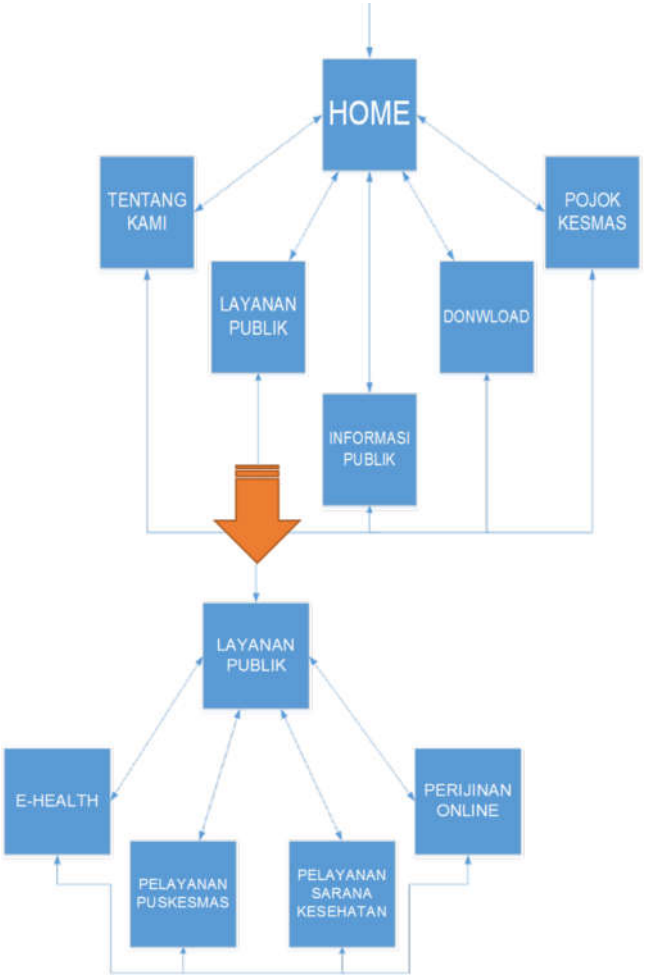


Fig. 3. The web structural navigation. The translation labels in English are the following: Tentang Kami (SHD Profile), Layanan Publik (Public Service), Informasi Publik (Public Information), Download (Health Data and Report), Pojok Kesmas (Community Health Center Corner), Pelayanan Puskesmas (Community Health Center Service), Pelayanan Sarana Kesehatan (Public Health Facility Service), and Perijinan Online (Online Health-related License Application)

F. Designing the Solution: Storyboard and Mock-Up

The storyboard visualizes the possible scenario of user interaction on each webpage. It also facilitates the exploration of user experience with the website. We developed the storyboard for each webpage based on the user requirement and the task analysis. An example of storyboard for homepage is shown if Fig. 4. The last step on designing the solution was building the website prototype based on the mock-up approach. Fig. 5. shows the example of homepage prototype.

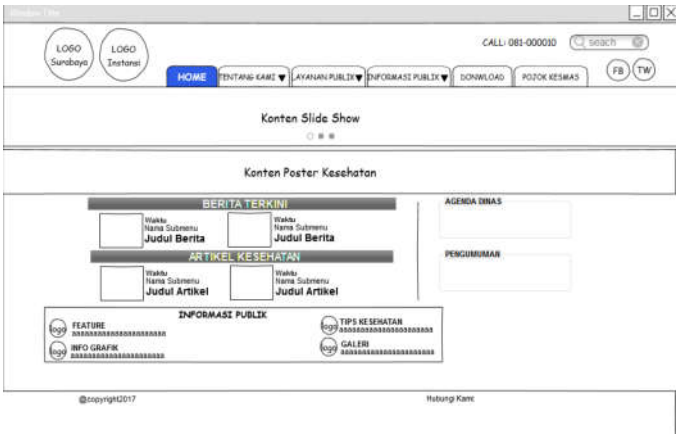


Fig. 4. Storyboard of SHD homepage website

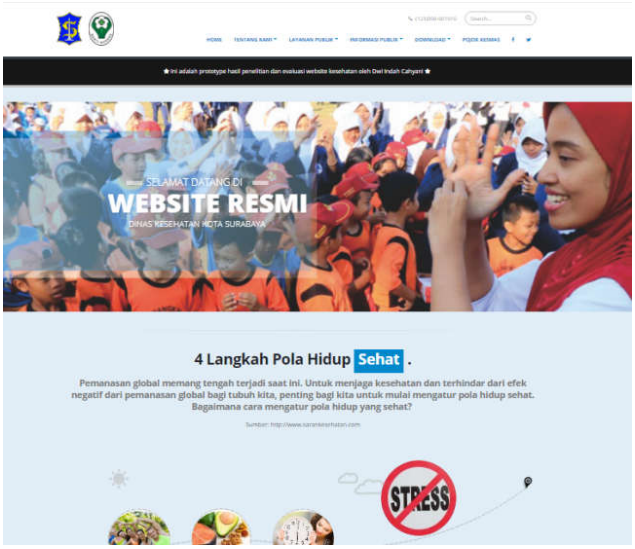


Fig. 5. The homepage mock-up (in Indonesian language)

G. *Evaluating the Solution Against the Requirements (Heuristic Evalution 2)*

We conducted heuristic evaluation 2 to evaluate the correctness and the completeness of the developed solution, i.e., the new website prototype. The heuristic evaluation 2 used the same questionnaire and the same participants as in heuristic evaluation 1. Table V shows the heuristic evaluation 2 results.

TABLE V. THE RESULT OF NEW WEBSITE EVALUATION (HEURISTIC EVALUATION 2)

Heuristic Indicator	Question Code	Tabulated Score	Interval Interpretation
Aesthetic and minimalist design	AE1	90%	Strongly agree
	AE2	85%	Strongly agree
	AE3	75%	Agree
Visibility of system status	VI1	85%	Strongly agree
	VI2	90%	Strongly agree
	VI3	90%	Strongly agree
	Vi4	80%	Strongly agree
	VI5	90%	Strongly agree
Match between system and the real world	MA1	85%	Strongly agree
	MA2	90%	Strongly agree
	MA3	90%	Strongly agree
Recognition rather than recall	RE1	85%	Strongly agree
	RE2	85%	Strongly agree
Consistency and standards	CO1	85%	Strongly agree
	CO2	85%	Strongly agree
	CO3	70%	Agree
Flexibility and efficient of use	FL1	95%	Strongly agree
	FL2	90%	Strongly agree
Use control and freedom	US1	90%	Strongly agree
	US2	90%	Strongly agree
Error prevention	ER1	85%	Strongly agree
	ER2	80%	Strongly agree
Help users recognize, dialogue, and recovers from errors	HE1	80%	Strongly agree
Help and documentation	HD1	90%	Strongly agree

The new website evaluation (heuristic evaluation 2) exhibits improvements in all heuristic indicators compared to the existing website evaluation (heuristic evaluation 1). Only one question in one indicator scores below 75%, i.e., the question about the content update consistency (OC3). It occurs because the new website is still in the development phase, thus the content is not updated frequently. These improvements confirm that the user-centered development can capture the user requirement more comprehensively, thus increases the likelihood of system success. One of the examples is the inclusion of specific requirement for person with color-blind and person with dyslexia.

While most of the results correspond positively to the purpose of this study, further iterations of user-centered design are necessary to complete the new redesigned website. Further evaluation should be conducted by involving larger participants and representatives from the government stakeholders.

IV. DISCUSSION

This paper proposes the application of UCD to redesign a public health agency website, i.e., the Surabaya Health Department in Surabaya city. The UCD ensures the website development incorporating all users' needs. This is essential because the Surabaya Health Department website delivers public health information to broad audiences. Error and mistake in health information delivery can cause fatal consequences. Another consideration in the website development is the involvement of experts and practitioners from multidiscipline fields to assure the delivery of accurate health information in the correct presentation.

V. Conclusion

This study applies the user-centered design to develop a website that can improve public health communication and health information exchange between a government agency (i.e., Surabaya Health Department) and the society of Surabaya. The results of this study support three main contributions, i.e., a detail identification of the targeted users, the inclusion of specific requirements for users with special needs, and the application of UCD to improve the user experience. The enhancement of the user experience is exhibited in the heuristic evaluation scores between the existing website and the new redesigned website.

References

[1] J. Choi and S. Bakken, "Web-based education for low-literate parents in Neonatal Intensive Care Unit: development of a website and heuristic evaluation and usability testing.," *Int. J. Med. Inform.*, vol. 79, no. 8, pp. 565–75, Aug. 2010.

[2] S. Kim, K.-H. Lee, H. Hwang, and S. Yoo, "Analysis of the factors influencing healthcare professionals' adoption of mobile electronic medical record (EMR) using the unified theory of acceptance and use of technology (UTAUT) in a tertiary hospital," *BMC Med. Inform. Decis. Mak.*, vol. 16, no. 1, p. 12, Dec. 2015.

[3] Y.-L. Chiu and C.-C. Tsai, "The roles of social factor and internet self-efficacy in nurses' web-based continuing learning.," *Nurse Educ. Today*, vol. 34, no. 3, pp. 446–50, Mar. 2014.

[4] M. Georgsson and N. Staggers, "Quantifying usability: an evaluation of a diabetes mHealth system on effectiveness, efficiency, and satisfaction metrics with associated user characteristics," *J. Am. Med. Informatics Assoc.*, vol. 23, no. 1, pp. 5–11, Jan. 2016.

[5] I. Puspitasari, "The impacts of consumer's health topic familiarity in seeking health information online," in *Proceedings - 2017 15th IEEE/ACIS International Conference on Software Engineering Research, Management and Applications, SERA 2017*, 2017.

[6] C. Abras, D. Maloney-Krichmar, and J. Preece, "User-centered design," *Bainbridge, W. Encycl. Human-Computer Interact. Thousand Oaks Sage Publ.*, 2004.

[7] D. A. Norman, "Human-centered design considered harmful," *interactions*, 2005.

[8] I. Puspitasari, "Stakeholder's expected value of Enterprise Architecture: An Enterprise Architecture solution based on stakeholder perspective," in *2016 IEEE/ACIS 14th International Conference on Software Engineering Research, Management and Applications, SERA 2016*, 2016.

[9] M. MAGUIRE, "Methods to support human-centred design," *Int. J. Hum. Comput. Stud.*, vol. 55, no. 4, pp. 587–634, Oct. 2001.

[10] R. Yáñez Gómez, D. Cascado Caballero, and J.-L. Sevillano, "Heuristic evaluation on mobile interfaces: a new checklist.," *ScientificWorldJournal.*, vol. 2014, p. 434326, Sep. 2014.

[11] M. I. Dahnil, K. M. Marzuki, J. Langgat, and N. F. Fabeil, "Factors Influencing SMEs Adoption of Social Media Marketing," *Procedia - Soc. Behav. Sci.*, vol. 148, pp. 119–126, Aug. 2014.

[12] I. Thoifah, *Statistika pendidikan dan metode penelitian kuantitatif*. Malang: Madani, 2015.

[13] A. Burrell and A. C. Sodan, "Web Interface Navigation Design: Which Style of Navigation-Link Menus Do Users Prefer?," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*, 2006, pp. 42–42.

[14] L. Rello and R. Baeza-Yates, "Good fonts for dyslexia," in *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '13*, 2013, pp. 1–8.

Determining The Senior High School Major Using Agglomerative Hierarchial Clustering Algorithm

Adhitya Nugraha
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
adhitya@dsn.dinus.ac.id

Mahendra Arista Harum Perdana
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
mahendraarista12@gmail.com

Heru Agus Santoso
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
heru.agus.santoso@dsn.dinus.ac.id

Junta Zeniarja
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
junta@dsn.dinus.ac.id

Ardytha Luthfiarta
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
ardytha.luthfiarta@dsn.dinus.ac.id

Ayu Pertiwi
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
ayu.pertiwi@dsn.dinus.ac.id

***Abstract*— Determining the senior high school major is still a dilemma for some junior high school students. The selection of high school majors must be tailored to the interests, talents and academic skills of students so that later students can develop a better competencies, attitudes and academic skills in the new environment. The selection of the appropriate high scholl major will influence students' interests and abilities in exploring a field of science so that later it will be easier for students to go to university which is expected and in accordance with their current interests and abilities. This will obviously be very beneficial for the student in preparing for his future.**

Clustering is one technique known in the data mining process. The core concept of clustering is to group a number of data or objects into a group or several groups where each group contains data that has similarities that are very close to other data. There are two types of grouping methods known as hierarchical clustering and partitioning. The hierarchical clustering method consists of several types, namely complete linkage clustering, single linkage clustering, average linkage clustering and centroid linkage clustering. While the partitioning method itself consists of the following types namely k-means clustering and k-means fuzzy clustering.

In this study, the authors have applied and analyzed the Agglomerative Hierarchical Clustering technique in the data of students of SMP Negeri 2 Purwodadi to classify students based on their respective interests and skills to fit the selection of high school majors. In the implementation, the author uses 5 attributes of pre-processing results which are used as experimental data processing variables. The results of this study succeeded in developing a prototype application that has implemented the Agglomerative Hierarchical Clustering algorithm which is used to visualize data processing so that it can help students determine high school majors. From the various experiments that have been carried out, this application has shown good resultsl.

Keywords—Data Mining, Clustering, Agglomerative Hierarchical Clustering Algorithm, Major Determination in Senior High School

I. INTRODUCTION

Selection of majors by the students according to interests when enrolling in high school has the aim to provide opportunities for students to develop competence attitudes and competence skills of students according with their interests, talents, and academic skills in a group of subjects in science. Determination of the majors will affect the next academic level and will affect the field of science or studies for students who will continue on to the university level so that the selection of incorrect majors could harm students and their future [1].

Placement of students according to ability or capacity is often called the selection of majors by students in senior high school are determined by academic ability that is supported by a factor of interest, because of the characteristics of a science requires the same characteristics of who will be studied. Thus, students who studied a science that correspond to the characteristics of their personality will feel comfortable when studying that science. Student interests can affect the quality of student achievement in a particular field of study. A student that interested in mathematics, for example, will more focus on mathematics than other students who are not interested in mathematics. Because the concentration of intensive attention to the material, students will study harder and try to achieve the achievement [2].

Decision making in choosing majors according with the curriculum in 2013 is determining the majors when students will be enrolled in senior high school by looking at several factors, including the score in progress report when in the junior high school, the score of junior high school national examination, recommendations from counseling teacher / counselor in junior high school. The programs that are available in the senior high school include Science (IPA), Social (IPS), and Linguistics (Bahasa)[1]. In addition, the selection of majors aims to adjust capabilities and interest of the students to the field chosen by the student. Selection of the appropriate majors will increase the interest and provide a sense of comfort to the students in learning. With the same basic capabilities expected in learning activities can proceed

smoothly, without any students who have difficulty in learning and increase the interest and student achievement. In contrast, lack of interest in learning as a result of a mistake in choosing majors [2].

Because too many students who will choose the majors, counseling teacher have difficult to perform a calculation to determine the majors because the process of selecting majors still done manually by grouping the student's progress report score based on the relevant courses, and still consider the student's interests and recommendations from counseling teacher, so that way is less effective and less efficient because they have extra work and requires a long time [2].

Clustering is one technique known in the data mining process. The core concept of clustering is to group a number of data or objects into a group or several groups where each group contains data that has similarities that are very close to other data. There are two types of grouping methods known as hierarchical clustering and partitioning. The hierarchical clustering method consists of several types, such as complete linkage clustering, single linkage clustering, average linkage clustering and centroid linkage clustering. While the partitioning method itself consists of the following types, such as k-means clustering and k-means fuzzy clustering. [3].

II. RELATED WORKS

Fuzzy C-Means (FCM) algorithm can provide recommendations or determination of the optimal majors for the students and also how Genetic algorithms can distribute students into classes based on the available quota. The process of grouping or the determination of these majors are obtained based on the evaluation value of each subject of the majors and based on the evaluation value of each subject related to the majors [4].

Clustering is the process of grouping a collection from the objects that have the same members into classes of the same object. The purpose of clustering is to group a set of data or objects that have a similarity in one cluster and different from the objects in other clusters. The method that used in this research is Agglomerative Hierarchical Clustering method. This method is done by grouping a set of objects (many objects into one piece). Begins by placing any object on clustering and then combine the clustering structure becomes larger clustering until all the objects currently on the same clustering or to fulfill predetermined conditions. This algorithm conducted with a bottom-up approach [5].

III. THEORETICAL BACKGROUND

Data mining is a method by finding a knowledge in a database that is large enough. Data mining is the process of digging and analyzing very large amounts of data to obtain something right, new, useful and ultimately understandable a pattern or patterns in the data [6].

Hierarchical Clustering Method (HCE) is one of the methods used in performing clustering [3]. Clustering that formed from HCE in the form of a hierarchical (tree). In hierarchial clustering, data is not divided into a clustering in one step but in a series of clustering. The division can be done from a cluster that containing all objects to some clustering in the number of n where each value of n contains an object [3].

Hierarchial Clustering method is classified into two types, namely Agglomerative and Divisible Hierarchial Clustering, depending on the composition conducted from the form of bottom-up or top-up [7]. The method used in this study is Agglomerative Hierarchical Clustering Method. This method is done by grouping a set of objects (many objects into one piece).

Begins by placing any object on clustering and then combine the clustering structure becomes larger clustering until all the objects that are in the same clustering or to fulfill predetermined conditions. This algorithm performed with a bottom-up approach [3].

The algorithm used in Hierarchial Clustering Method is Hierarchical Clustering Algorithm. In the algorithm, if given a number N, the object will be clustering with proximity matrix N * N, then [9]:

1) Positioning Figures Beginning with making each object into one clustering. If the available number N, the object to be in clustering, it will form a clustering as many N which contain only one object.
2) Find the nearest of the pair clustering and combined into one clustering. Clustering can be combined into one clustering if it has an object with the closest similarity to the object on the other clustering.
3) Calculate the distance or similarity between the newly formed clustering and other clustering that has existed before.
4) Repeat steps 2 and 3 until all objects that performed clustering process currently on one clustering

Agglomerative Hierarchical Clustering is a method by using the strategy design Bottom-Up which begins by placing any object as a cluster itself (atomic cluster) and then combine atomic clusters into a cluster, which is bigger and bigger until finally all the objects fuse into a cluster or the process can also be stopped if it has reached the limits of specific conditions [3].

The measure of distance that is used to combine two objects cluster is Euclidean Distance [3], which can be seen in the following equation:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2} \quad (1)$$

IV. EXPERIMENT AND RESULT

This research uses data that taken from SMP Negeri 2 Purwodadi. Data that provided by the school was presented in hard copy. Student data that provided by SMP Negeri 2 Purwodadi is data about student name and student's national exam score. This dataset consist of 500 records data, with 10 attributes, and with 2 different generations such as: IPA and IPS.

The original student data given by SMPN 2 Purwodadi can't directly use as the experimental data. It is because of the original student data does not have enough predictor variable

attributes that become the requirements of supervised learning in this study.Detail of data are shown on table I.

TABLE I. STUDENT RAW DATA

No		NAMA	TUJUAN	JENIS	UN				US						Jurusan
Urut	Pendaf			KELAMIN	BI	Bing	Mtk	IPA	Agm	PKN	IPS	BI	Seni	TIK	
1	305	ABDULLOH ZAKARIA		L	72	60	32,5	55	92	80	82,5	80	81,7	83,3	IPS
2	523	ABDUR ROCHMAN		L	72	52	35	60	88	87	85	87,9	88,5	87	IPS
3	512	ACHMAD TAUFIQ H		L	78	40	50	35	87,7	85	82,6	81,6	80	93	IPA
4	422	AENI APRILIANTI		P	64	76	60	60	93	93,2	83	83,4	87	88	IPA
5	446	BAROYAH		P	88	78	75	82	91,5	87	85	84,3	91,8	85,5	IPS
6	42	DENI KARTIKA SARI		P	60	82	57,5	45	92,5	80	85	85	85	80	IPA
7	486	DERINTAN ARUS SUKATI		P	82	36	42,5	40	90	78	77,5	82,5	80	83,3	IPA
8	517	DETY KOMARIYAH		P	70	70	45	62,5	79,2	76,7	73,3	75,8	76	85	IPS
9	171	DEWI ATIKA		P	84	36	47,5	57,5	84	83	87	82,49	88	87	IPA
10	290	DEWI KHUMAEDAH		P	70	56	35	47,5	84,6	81,7	82,1	82	86,7	83,3	IPA
11	169	EGA UTAMI		P	72	64	40	45	96	96,3	82	84,3	87	91,5	IPA
12	303	FALA AULIA FADHILAH		P	72	60	45	47	92	89	80	83	78	78	IPA
13	427	FAZHIA ULLY DIANTI		P	72	34	42,5	42,5	88	87	77,5	85	80	80	IPA
14	555	FITRIYA		P	64	52	35	42,5	80,9	83,5	80,8	81	90	88,3	IPS
15	469	HIDAYATUS SIBYAN		P	66	48	35	70	94	92,3	92,6	92,2	88	84	IPS
16	350	IIS DAHLIA		P	70	54	42,5	50	90	86	85	72,5	84	83	IPA
17	27	IKA VIDIARTI		P	68	46	52,5	70	86,6	81,5	82,5	87	77	84,6	IPA
18	473	INDAH NAVITASARI		P	72	58	52,5	52,5	90,7	89,1	88	89,3	90	90	IPA
19	324	KARINA SAHRANI		P	77	67	55	69	87	86	90	74	92	78	IPA
20	83	KARYUTI		P	84	52	37	40	80	80	87,5	81,5	80	82	IPS

To perform experimental testing in this study, there are some differences in data processing/preprocessing phase. It means the data preparation need to be performed in the very beginning of the experimental process before doing a testing. There are some steps taken to get the original student data become an example set used in this study. Those are data selection and data reduction.

TABLE II. EXAMPLES OF DATASET

UN BI	UN Bing	UN Mtk	UN IPA	Label
72	60	32,5	55	IPS
72	52	35	60	IPS
78	40	50	35	IPA
64	76	60	60	IPA
88	78	75	82	IPS
60	82	57,5	45	IPA
82	36	42,5	40	IPA
70	70	45	62,5	IPS
70	56	35	47,5	IPA
72	34	42,5	42,5	IPA
64	52	35	42,5	IPA
66	48	35	70	IPA
70	54	42,5	50	IPA
72	58	52,5	52,5	IPS
84	52	37	40	IPS
68	68	22,5	40	IPA
80	54	45	47,5	IPA
74	68	57,5	65	IPA
78	74	90	62,5	IPA
60	70	30	57,5	IPS
66	38	30	52,5	IPS

Table II explained that data transformation is performed by changing some of data to make data mining process become easier after the dataset are selected. Data reduction process is the process performed to clean the data record in student data that does not selected. Moreover, this step is also performed to deleting some attributes that does not have the integrity for information retrieval with data mining process.

From these regulation, cleaning the data in student data from unselected student data record. After the selection process, not all student data is used because of the limitation variables from some students. The student that does not have a valid status information in predictor variable must be deleted. Deleting some attributes in customer data that does not have an integrity for retrieval information. Those attributes are No Urut, No Pendaf, Nama, Tujuan Sekolah and Jenis Kelamin. Those attributes can be deleted because it does not give any information in data mining process. Moreover, those attributes are used only as an information to the school, which is SMP Negeri 2 Purwodadi, about the student. After data reduction process, student data used for this study is ready to use. This data also called as example set. It consists of 290 records of student data with 5 attributes. Those attributes are UN BI, UN Bing, UN MTK, UN IPA, and Jurusan.

After the example dataset ready to be proceed, the testing phase is started. The best testing result shown at the figure below.

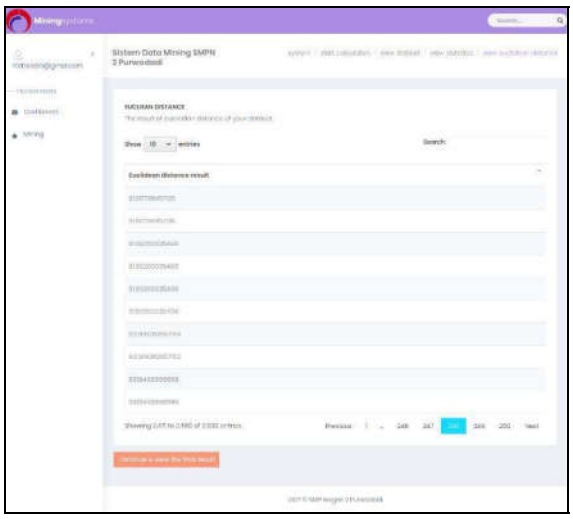


Fig. 1. Euclidean Distane on the system

Figure 1 show that the Euclidean Distance that serves the result of Euclidean distance calculation of the data set. There are 2500 entries of Euclidean distance result from the data set in previous step. The user can click “continue & view the final result” button to know the final result all of the calculation process.

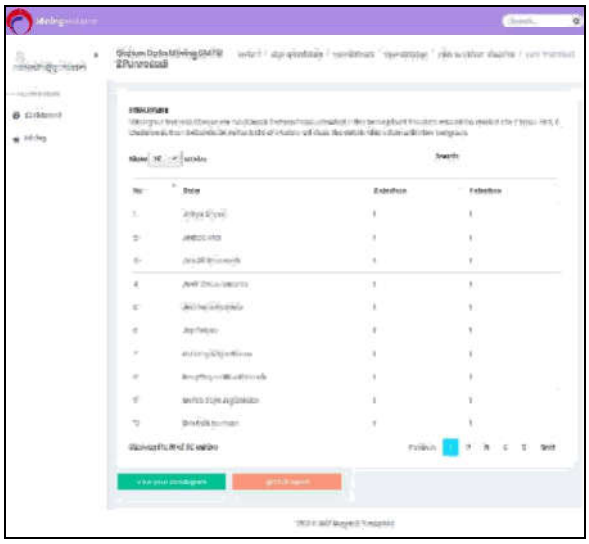


Fig. 2. Final Result of Clustering

Figure 2 shows that the Final Result of Clustering that contain final result of whole the calculation process based on the data set that have been uploaded in the form upload. The data here will be divided into 2 types. First 2 clusters and then 3 clusters. So, in the table of clusters will show the detail which data will follow the group.

V. EXPERIMENT AND RESULT

In this study, the implementation of Agglomerative Hierarchical Clustering Algorithm for clustering prediction to determine the major in senior high school get the fair result. The best testing result shown 277 students on cluster 1 and 13 students on cluster 2 from 290 record student data. Agglomerative Hierarchical Clustering Algorithm is the one method that can be applied to determine the major in senior high school that is used to help SMPN 2 Purwodadi student.

REFERENCES

[1] M. Nur, "Peminatan pada Pendidikan Menengah," Jakarta : Menteri Pendidikan dan Kebudayaan Republik Indonesia, 2014.

[2] Bahar, "Penentuan Jurusan Sekolah Menengah Atas dengan Algoritma Fuzzy C-Means," Thesis Program Pasca Sarjana Udinus, Semarang, 2011.

[3] J. Han and K. Micheline, Data Mining Concepts and Technique, Morgan Kaufmann, 2001.

[4] Fajar Akbar, "Clustering Data Nilai Siswa SMA Untuk Penentuan Jurusan Menggunakan Algoritma Fuzzy C-Means," 2015.

[5] Marjiyono, " Penerapan Algoritma AHC Algorithm Dalam Aplikasi Pembagian Kelas Siswa Baru," 2015.

[6] J. Han and M. Kamber, Data Mining: Concepts and Techniques Second Edition, San Francisco: Morgan Kaufmann Publishers, 2006.

[7] R. Margo dan E. Ari Eka, Penerapan Algoritma Clustering Untuk Mengelompokkan Ketertarikan Siswa Berdasarkan Aktivitas di Model Pembelajaran Elektronik, 2014.

Influence of Word Normalization and Chi-squared Feature Selection on Support Vector Machine (SVM) Text Classification

Ardy Wibowo Haryanto[1], Edy Kholid Mawardi[2], Muljono[3]
[1,2,3]Informatics Engineering Department
Dian Nuswantoro University
Semarang, Indonesia
[1]ardy.dinus@gmail.com, [2]edycoleee@gmail.com, [3]muljono@dsn.dinus.ac.id

Abstract*—In this study, we used SVM for text classification. There is stemming or lemmatization word normalization with the addition of Chi-square feature selection on the classification that we made. There are also pre-processing data being performed, namely stopwords removal and tokenize. We used BBC dataset containing 2,225 documents and 5 categories. There are 21,813 features resulting from the use of stemming and 31,007 features resulting from the use of lemmatization. Each feature represents the number of words that come out in the document. We used confusion matrix to evaluate the results of text clasification. SVM text classification performance using stemming enhanced by Chi-squared (method 1) get better results than using lemmatization enhanced by Chi-squared (method 2). The best performance was obtained using 80% feature reduction where method 1 received a precision value of 95%, a recall value of 95%, and an accuracy value of 95.05%. Method 2 only received a precision value of 93%, a recall value of 93%, and an accuracy value of 93.24% using the same amount of feature reduction.

Keywords— text classification; SVM; Chi-squared; stemming; lemmatization; machine learning

I. INTRODUCTION

Information Retrieval (IR) is the science to find information that can be done on various forms of data such as text, image, and sound. There are various IR applications such as in the extraction of information on the text before the text classification is done. Text classification is the process of automatically classifying unknown text by suggesting the label that is most likely to belong to it. Some text classification applications are news groupings, document organization, and spam email filtering. Some of the methods that can be used for text classification are k-Nearest Neighbor (KNN), Naive Bayes, Decision Tree, and SVM [1], [2].

Toman et al used the Naive Bayes for text classification [3]. Their studies focus on proving and comparing the effects of word normalization. They compared algorithms related to stemming and lemmatization, namely Lovins and Iterated Lovins, Paice, Porter's stemmer, EWN-based lemmatization with and without using indexes. The result is the use of word normalization has no significant effect in improving the performance of text classification. The best pre-processing approach to text classification in these studies is to implement stopwords removal only and eliminate word normalization.

Further studies were conducted by A. Sarkar et al. They prefer to use SVM to do text classification [4]. There are two stages of pre-prosessing data that they did, change to lowercase and stopwords removal. There are only 2 categories in their text classification. They claimed that their text classification is successfully implemented and got a reasonable rate of accuracy.

SVM is also used by Yin et al with slight modifications [5]. First they got the model outcome of the training using SVM. Then, the model is used to label unlabeled documents. The result of labeling the document is used for retraining to get the final model. They only use stopwords removal for pre-processing data. Their method successfully improves the performance of SVM to classify text.

Another studies were conducted by S. D. Sarkar et al [6]. They used the Naive Bayes for text classification. In their studies, they developed the FS-CHICLUST feature selection to improve the poor performance of Naive Bayes for text classification. The FS-CHICLUST feature selection is an enhancement of Chi-squared feature selection. Change to lowercase, stopwords removal, and stemming done for pre-processing data and TF-IDF selected for feature extraction.

In these studies, we tried to combine the use of stemming and lemmatization done by Toman et al with the addition of Chi-squared feature selection without modification which has been done by S. D. Sarkar et al [3], [6]. We used the SVM for text classification. The outline of this paper is as follows. In section III we convey the studies that have been conducted. Then in section IV, we describe the dataset and the method. After that in section V, we describe the results. There are conclusions in section VI.

II. THEORIES

A. Support Vector Machine (SVM)

The SVM is a supervised learning which uses the statistical approach. SVM is mainly used to solve regression and classification problems. SVM was originally proposed to solve binary classification [7]. SVM has the basic idea of implicitly mapping training data into a high-dimensional feature space. A hyperplane is constructed in this feature space that maximizes the separation margin between the hyperplane and the points located closest to it as a supporting vector. Kernel function is a

mapping function that mapping the input space into a high-dimensional feature space. The most commonly used kernel functions are the Polynomial, Sigmoid, and Radial Basis Function (RBF) [8].

B. Chi-squared

Chi-squares is generally used to measure the lack of independence between term (t) and class or category (c) and compared to the X^2 distribution with one degree of freedom [6], [9]. Chi-squares is shown in the formula:

$$X_{(c,t)} = \frac{D(AJ-BI)^2}{(A+I)(B+J)(A+B)(I+J)} \quad (1)$$

D is the number of documents. A is the number of documents class c containing t. B is the number of documents containing t occuring without c. I is the number of documents class c occuring without t. J is the number of documents of other classes without t.

C. Stemming and lemmatization

Both stemming and lemmatization produce a normal form with some differences. Stemming only produces an approximate form of a word. Lemmatization removes the suffix of a word completely to get the basic word form or replaces the suffix of a word with a different one.

III. Previous Studies

Toman et al conducted studies focusing on the proof and comparison of the use of word normalization before being classified using Naive Bayes [3]. They used two datasets that have different languages are English and Czech. For English dataset, they used datasets from Reuters Corpus Volume 1. The dataset contained 8,000 documents that have 6 different categories. For Czech dataset, they build it from documents belonging to the Czech News Agency. The dataset contained 8,000 documents that have 5 different categories. They did not explain the division of datasets for training and testing. They did pre-processing data by doing stopwords removal and word normalization. Six stemming and lematization algorithms for word normalization that they used are Lovins and Iterated Lovins, Paice, Porter's stemmer, EWN-based lemmatization with and without using indexes.

A. Sarkar et al used the SVM for news classification [4]. Their dataset consists of 500 titles of books with 250 related to computer science and 250 related to biology. For each category, they divided them into 200 documents for training and 50 documents for testing. They did pre-processing data by doing converting into lower case and stopwords removal.

Yin et al used the SVM with slight modifications for news classification [5]. They build dataset from documents belonging to the Sina weibo and Tencent weibo. The dataset contained 56,985 shot texts that have 5 different categories. The distribution of short texts per category for training and testing is shown in Table I.

TABLE I. DISTRIBUTION OF SHOT TEXTS PER CATEGORY

	Training	Testing	Total
Politics	7973	3417	11390
Economy	7981	3421	11402
Education	7968	3415	11383
Entertainment	7990	3424	11414
S & T	7977	3419	11396

They only use stopwords removal for pre-processing data.

The last are studies that have been done by S. D. Sarkar et al [6]. They conducted studies focusing on developing FS-CHICLUST feature selection to improve the poor performance of Naive Bayes for text classification. They used 13 different datasets that have different numbers of documents and categories. For each dataset, they divided it by 70% of documents for training and 30% of documents for testing. They did pre-processing data by doing converting into lower case, stopwords removal, and stemming. They used TF-IDF for feature extraction. Thet also used Chi-squared and their FS-CHICLUST for feature selection.

IV. Proposed Method

A. Dataset

We used BBC dataset in our studies. The documents on this dataset come from the BBC news website corresponding to stories in five topical areas from 2004-2005. This dataset consists of 2,225 documents that are divided into 510 documents categorized as business, 386 documents categorized as entertainment, 417 documents categorized as politics, 511 documents categorized as sport, and 401 documents categorized as tech.

There are pre-processing data given to that dataset, namely stopwords removal, tokenize, stemming or lemmatization. The collection of words from all existing documents that have been given pre-processing became a feature used in this study. Each feature contains the number of word that come out on a document. There are 21,813 features for data that is preprocessed using stemming and 31,007 features for data that is preprocessed using lemmatization.

B. Method

The purpose of our studies is to compare the performance of SVM using the same dataset with different pre-processing data accompanied by the addition of Chi-squared feature selection in classifying the news. The system to be made is shown in Fig. 1.

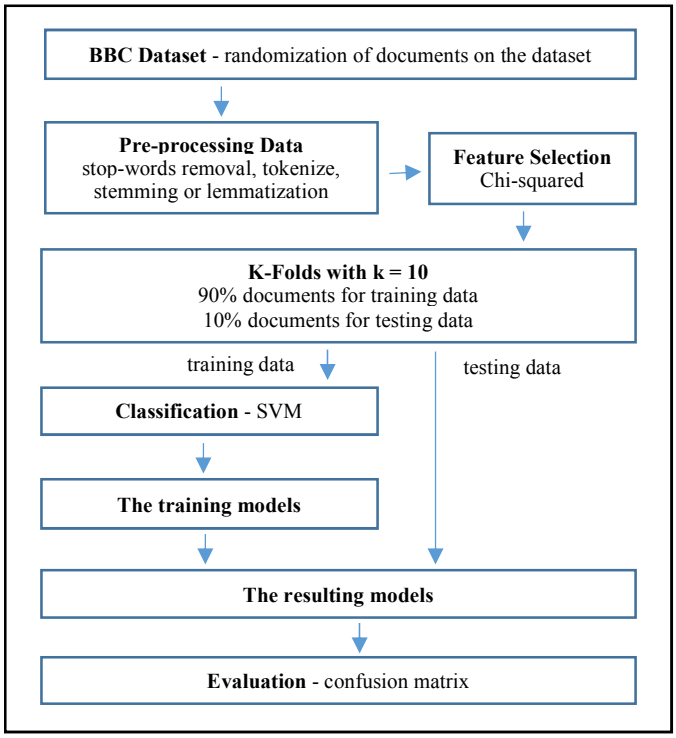


Fig. 1. The classification flow

Documents on the BBC dataset are randomized before going into pre-processing data. The difference of pre-processing data that we want to compare only in word normalization, using stemming or lemmatization. We used Lancaster Stemmer for stemming and WordNet Lemmatizer for lemmatization. Both were given Chi-square feature selection with a feature reduction amount of 20% which gradually increased up to 80%. Using k-Folds with k = 10, we get 90% of documents for training data and 10% of documents for testing data. We did the training using SVM with RBF kernel to produce models. We tested the models that have been produced and evaluated using confusion matrix.

V. RESULTS

The confusion matrix is one commonly used to evaluate text classification results [10], [11], [12], [13]. The confusion matrix is shown in Table II.

TABLE II. THE CONFUSION MATRIX

	Predicted: Yes	Predicted: No
Actual: Yes	True Positive (TP)	False Negative (FN)
Actual: No	False Positive (FP)	True Negative (TN)

The confusion matrix can be used for binary and multi-class classification evaluation. Some of the metrics listed in Table II will be extended if used for multi-class classification. From the confusion matrix, we can use the formula:

$$precision = \frac{TP}{TP+FP} \quad (2)$$

$$recall = \frac{TP}{TP+FN} \quad (3)$$

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Precision is used to measure the positive patterns that are correctly classified from the total classified patterns in a positive class. Recall is used to measure the effectiveness of a classifier to classify positive patterns. Accuracy is used to measure the overall effectiveness of a classifier.

We tested and evaluated the models resulting from the training process. We also compared the evaluation results. The confusion matrix for news classification produced using SVM with stemming and Chi-squared feature selection is shown in Table III to Table VII.

TABLE III. STEMMING - FEATURE REDUCTION OF 0%

Actual / Predict	enter tainment	sport	business	tech	politics
entertainment	25	10	3	0	3
Sport	0	47	0	0	0
Business	0	0	57	1	1
Tech	0	1	8	19	0
Politics	1	3	11	0	32

TABLE IV. STEMMING - FEATURE REDUCTION OF 20%

Actual / Predict	enter tainment	sport	business	tech	politics
entertainment	32	5	2	0	2
Sport	0	47	0	0	0
Business	0	0	57	1	1
Tech	0	1	5	22	0
Politics	1	2	10	0	34

TABLE V. STEMMING - FEATURE REDUCTION OF 40%

Actual / Predict	enter tainment	sport	business	tech	politics
entertainment	38	2	0	0	1
Sport	0	47	0	0	0
Business	0	0	57	1	1
Tech	0	1	3	24	0
Politics	2	1	9	0	35

TABLE VI. STEMMING - FEATURE REDUCTION OF 60%

Actual / Predict	enter tainment	sport	business	tech	politics
entertainment	39	1	0	0	1
Sport	0	47	0	0	0
Business	0	0	57	1	1
Tech	0	1	3	24	0
Politics	2	0	5	0	40

TABLE VII. STEMMING - FEATURE REDUCTION OF 80%

Actual / Predict	enter tainment	sport	business	tech	politics
entertainment	40	0	0	1	0
Sport	0	47	0	0	0
Business	1	0	56	1	1
Tech	0	1	0	27	0
Politics	1	0	4	1	41

From Table III to Table VII it can be seen that there is a trend of performance improvement. The precision, recall and accuracy of news classification produced using SVM with stemming and Chi-squared feature selection are shown in Table VIII.

TABLE VIII. THE PRECISION, RECALL, AND ACCURAY FOR STEMMING

Feature Reduction	Precision Avg	Recall Avg	Accuracy
0%	84%	81%	81.08%
20%	88%	86%	86.49%
40%	91%	91%	90.54%
60%	93%	93%	93.24%
80%	95%	95%	95.05%

From Table VIII, it can be seen that the average of precision, the average of recall, and accuracy will increase along with the enlargement of feature reduction. Based on this, it can be said that Chi-squared feature selection can improve the performance of text classification by using SVM and stemming. The confusion matrix for news classification produced using SVM with lemmatization and Chi-squared feature selection is shown in Table IX to Table XIII.

TABLE IX. LEMMATIZATION - FEATURE REDUCTION OF 0%

Actual / Predict	enter tainment	sport	business	tech	politics
entertainment	13	17	8	1	2
sport	0	47	0	0	0
business	0	2	53	3	1
tech	0	2	11	14	1
politics	1	3	13	0	32

TABLE X. LEMMATIZATION - FEATURE REDUCTION OF 20%

Actual / Predict	enter tainment	sport	business	tech	politics
entertainment	21	12	5	1	2
sport	0	47	0	0	0
business	0	2	54	2	1
tech	0	1	11	16	0
politics	1	3	12	0	31

TABLE XI. LEMMATIZATION - FEATURE REDUCTION OF 40%

Actual / Predict	enter tainment	sport	business	tech	politics
entertainment	28	8	2	1	2
sport	0	47	0	0	0
business	0	0	57	1	1
tech	1	1	8	18	0
politics	1	3	9	0	34

TABLE XII. LEMMATIZATION - FEATURE REDUCTION OF 60%

Actual / Predict	enter tainment	sport	business	tech	politics
entertainment	35	3	0	1	2
sport	0	47	0	0	0
business	0	0	57	1	1
tech	0	1	5	22	0
politics	2	1	7	0	37

TABLE XIII. LEMMATIZATION - FEATURE REDUCTION OF 80%

Actual / Predict	enter tainment	sport	business	tech	politics
entertainment	40	0	0	0	1
Sport	0	47	0	0	0
Business	1	0	56	1	1
Tech	1	1	2	24	0
Politics	2	0	4	1	40

From Table IX to Table XIII it can be seen that there is a trend of performance improvement. The precision, recall and accuracy of news classification produced using SVM with lemmatization and Chi-squared feature selection are shown in Table XIV.

TABLE XIV. THE PRECISION, RECALL, AND ACCURAY FOR LEMMATIZATION

Feature Reduction	Precision Avg	Recall Avg	Accuracy
0%	76%	71%	70.72%
20%	80%	76%	76.13%
40%	85%	83%	82.88%
60%	90%	89%	89.19%
80%	93%	93%	93.24%

From Table XIV, it can be seen that the average of precision, the average of recall, and accuracy will increase along with the enlargement of feature reduction. Based on this, it can be said that Chi-squared feature selection can also improve the performance of text classification by using SVM and lemmatization. Comparison of performance improvement from the use of stemming and lemmatization is shown in Fig. 2.

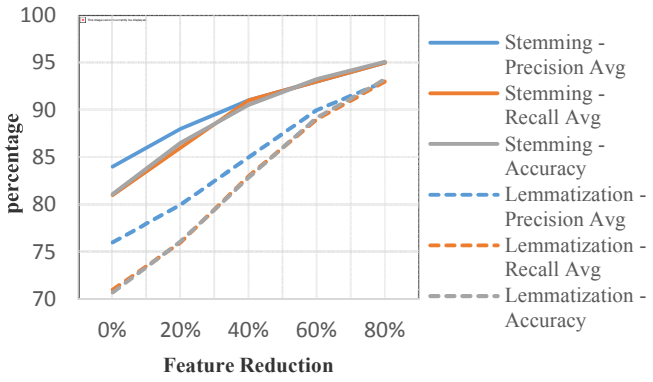


Fig. 2. The classification flow

From Fig. 2, it can be seen that the use of Chi-squared is more influential on the improvement of the performance of text classification by using SVM and lemmatization. However, the use of stemming and Chi-squared feature selection still yield better performance for text classification by using SVM.

VI. CONCLUSION

The use of Chi-square feature selection on the results of pre-processing data both with stemming and lemmatization can improve the performance of text classification using SVM. SVM text classification performance using stemming and Chi-square feature selection is superior to SVM text classification performance using lemmatization and Chi-square feature selection. The best performance was obtained using 80% feature reduction where SVM text classification using stemming enhanced by Chi-squared received a precision value of 95%, a recall value of 95%, and an accuracy value of 95.05%. SVM text classification using lemmatization enhanced by Chi-squared only received a precision value of 93%, a recall value of 93%, and an accuracy value of 93.24% using the same amount of feature reduction.

For future development, we would like to exploit the use of the modified SVM algorithm and FS-CHICLUST feature selection to classify the text. Studies can still be developed by deepening the use of pre-processing data, word normalization, feature selection, and other text classification algorithms. It can also be done with the use of feature extraction.

REFERENCES

[1] V. Bijalwan, V. Kumar, P. Kumari, J. Pascual, "KNN Based Machine Learning Approach for Text and Document Mining", International Journal of Database Theory and Application, vol. 7, no. 1, pp. 61-70, 2014.

[2] S. B. Mangal and V. Goyal, "Text News Classification System using Naïve Bayes Classifier", Intenational Journal of Engineering Sciences, vol. 3, 2014.

[3] M. Toman, R. Tesar, K. Jezek, "Influence of Word Normalization on Text Classification", INSCIT, 2006.

[4] A. Sarkar, S. Chatterjee, W. Das, D. Datta, "Text Classification using Support Vector Machine", International Journal of Engineering Science Invention, vol. 4, iss. 11, pp. 33-37, November 2015.

[5] C. Yin, J. Xiang, H. Zhang, J. Wang, "A New SVM Method for Short Text Classification Based on Semi-Supervised Learning", 4th International Conference on Advanced Information Technology and Sensor Application, 2015.

[6] S. D. Sarkar, S. Goswami, A. Agarwal, J. Aktar, "A Novel Feature Selection Technique for Text Classification Using Nave Bayes", International Scholarly Research Notices, vol. 2014, Published by Hindawi Publishing Corporation, 2014.

[7] P. V. Arivoli and T. Chakravarthy, "Document Clasification Using Machine Learning Algorithms - A Review", International Journal of Scientific Engineering and Research (IJSER), vol. 5, iss. 2, pp. 48-54, February 2017.

[8] M. H. Afif and A. R. Hedar, "Data Classification using Support Vector Machine Integrated with Scatter Search Method", Japan-Egypt Conference on Electronics, Communications, and Computers, 2012.

[9] C. D. Manning, R. Prabhakar, H. Schutze, "Introduction to Information Retrieval", vol. 1, Cambridge University Press, 2008.

[10] M. Hossin and M. N. Sulaiman, "A Review on Evaluation Metrics for Data Classification Evaluations", International Journal of Data Mining & Knowledge Management Process (IJDKP), vol. 5, no. 2, March 2015.

[11] M. Sokolova and G. Lapalme, "A Systematic Analysis of Performance Measures for Classification Tasks", Information Processing and Management 45, pp. 427-437, Published by Elsevier Ltd, 2009.

[12] Muljono, N. A. S. Winarsih, C. Supriyanto, "Evaluation of Classification Methods for Indonesian Text Emotion Detection", International Seminar on Application for Technology of Information and Communication, 2016.

[13] Muljono, U. Afini, C. Supriyanto, "Morphology Analysis for Hidden Markov Model based Indonesian Part-of-Speech Tagger", 1st International Conference on Informatics and Computational Sciences (ICICoS), 2017.

Sentiment Analysis using Weighted Emoticons and SentiWordNet for Indonesian Language

Nur Maulidiah Elfajr, Riyananto Sarno
Department of Informatics, Faculty of Information and Communication Technology
Sepuluh Nopember Institute of Technology, Surabaya, Indonesia
nur.maulidiah14@mhs.if.its.ac.id, riyananto@if.its.ac.id

***Abstract*- The large number of internet users caused increasing the number of social media users. Twitter is one of social media that have a large number of users in Indonesia. As a social media, twitter allows users to share information via status in a tweet. Due to the limitations of the use of text is only 280 characters, emoticons are commonly used in tweet. Emoticon can explain the condition or feeling which is described in a text-shaped punctuation mark. This paper will focus on creating emoticon dictionary and weighting of an emoticon. Emoticon dictionary contains a list of 384 emoticons describing a variety of feelings and emotions. The used dataset contains Indonesian language tweets from twitter API. We tried to analyze sentiment on existing datasets with reference scores in SentiWordNet. Weighting emoticons done under the assumption that the emoticons have more effect in a sentence than ordinary words. After that, we classify the results into three classes, namely sentiment positive, negative and neutral. We compared the results between the emoticon-based algorithm and without considering emoticons algorithm. Accuracy obtained on the emoticon-based using algorithm is 0.74.**

Keywords- Emoticons; Indonesian; Sentiment Analysis; SentiWordNet.

I. INTRODUCTION

Technological developments in the world growing rapidly affects the increase internet users. According to the 2017 Index of Tetra Pak launched last year, there are about 132 million internet users in Indonesia. Half of that number or 40% are users of social media and can be categorized as an addict. We can find users of social media in everyday life. Social media allow users to share information, share a moment or share their condition and feelings. In Indonesia, Twitter and Facebook social media are the most widely used by the community.

Twitter is one of social media that have text limit only 280 characters. So, users who want to share information in the form of status or tweet must ensure the status that they post are not more than 280 characters.

Therefore, the use of acronyms and the use of emoticons is very common in the Twitter status. Acronyms is the use of abbreviations in the text. While emoticons can describe the feeling or emotion's user without detailed explanation.

In addition, the use of emoticons in tweet can describes directly emotion's user. So it can easily identify sentiment of the tweet.

II. RELATED WORK

There are many studies have been conducted to analyze the sentiment by considering the composition of Lexical [1]. SentiWordNet as one of the common used lexicon in several studies [2] [3]. The approach used in SentiWordNet produce better accuracy than others [4].

In some experiments, the existence of the acronyms and emoticons are not considered. So if there is a sentence that contains acronym and emoticon, the sentence would not be valued. On paper [5] they tried to insert emoticons in the calculation to determine the sentiment. [6] create a dictionary of acronyms and classification for emoticons. However, the calculation is still not optimal, the emoticon will be considered if the condition of sentence is neutral. So it can reduce the number of neutral sentiment.

In this paper we combine some of the elements above. We're trying to build a lexical use SentiWordNet and take into consideration the existence of emoticons on t Indonesian language tweets. We will use the emoticons in the calculation so it will have a sentiment score on SentiWordNet.

SentiWordNet has limitation of language that is only available in English. Paper [6] tried implemented SentiWordNet into multilingual.

In the paper [7], they try to develop them using Indonesian. They interpret data from Indonesian into English so that it can be analyzed using SentiWordNet. To translate a specific language into the Indonesian language, paper [8] uses Bing as a media translator launched by Microsoft. While this paper will use Google translate translator which is considered pretty good at translating words and sentences. Detail of this method has been depicted in a fig.1 the proposed method.

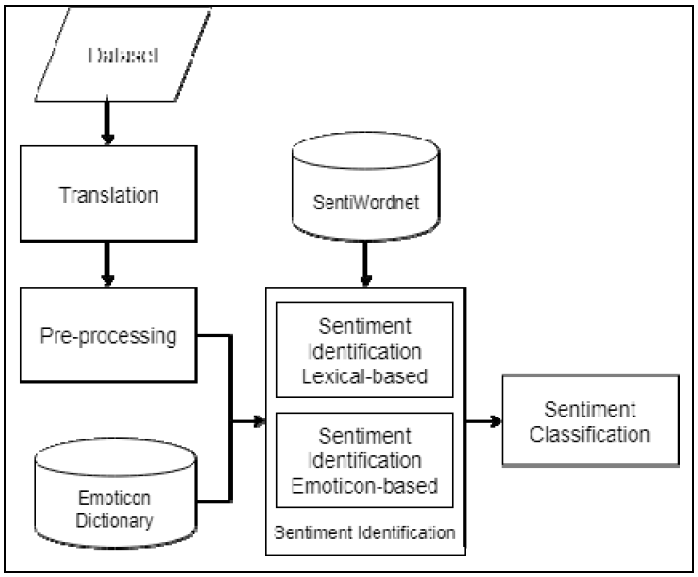


Fig. 1. Proposed Method

III. THE PROPOSED METHOD

A. Dataset

The dataset used data tweet Indonesian language. This data was taken with scrapling method using the twitter API called tweepy. Data tweet consists of id_user and tweet content. We labelling the data with the sentiment. The sentiments for each sentence are three classes, the which are positive, neutral, and negative. The dataset contains 100 sentences, the which are 32 positive class, 42 negative class, and 26 neutral class.

1. Example of positive sentence : Akhirnya selesai :))
2. Example of neutral sentence : Hari ini masuk sekolah :)
3. Example of negative sentence : Sedihnya ditinggal teman sekelas :(

B. Emoticon Dictionary

Emoticon is an interpretation of someone's feeling or emotion in text form icon. On each icon consists of ASCII and Unicode characters. Basically, emoticon consists of two types there are western and eastern.

1.) Western

Western emoticons type widely used in America and Europe. Western emoticons using all parts of the face are the eyes, the nose (not always) and mouth. emoticon western species are mostly written from left to right and then rotated 90 degrees in the direction of clockwise. Example emoticons:

:), :-(, : /, > :-), > _>, : '(

2.) Eastern

Western emoticons are widely used in Asia, especially Japan. Weastern emoticon type using eye as the main object. Eastern emoticons generally do not need rotation as in western

Example emoticons:

(^ _ ^;) (> _ <), ^ _ ^, (~ _ ~) , (* '∇`*)

We register all existing emoticon. according to Wikipedia sources [10] into list. We also consider it from paper [11] and paper [12]. This data consists of 384 eastern and western emoticons. From emoticons, we group similar emoticons depicting the moods or emotions together into one word. Thus forming an emoticon dictionary containing several words with each other's emotions. This grouping produce 46 emotions. Every emotion has its own list of members emoticon. For more details, see the Table. I

TABLE I. EMOTICON DICTIONARY

No.	Emotion	Emoticons
1.	Happy	':-)', ':)', ': -]', ':]', ': -3', ': 3', ': ->', ':>', '8-)', '8)', ': -}', ':}', ': o)', ': c)', ': ^)', '=]', '=)', '(? ^ o ^ ?) ',' (^ v ^) ',' (^ u ^) ',' (^? ^) ',' (^) o (^) ',' (^ O ^) ',' (^ o ^) ',' (^? ^) ',') ^ o ^ (',' (* ^? ^ *) ',' (????) '
2.	Laugh	'-D', ': D', '8D', '8D', 'xD', 'xD', 'XD', 'XD', '= D', '= 3' , 'B ^ D', '(* ^^) v', '(^^) v', '(^ _ ^) v', '(' - '*)', '(^ v ^)', '(^? ^)', '(·? ·)', '('? ')', '(???)'
3.	Excited	'\ (~ O ~) /', '\ (^ o ^) /', '\ (- o -) /', '? (^? ^)?', '? (^ O ^)?' '(* ^ 0 ^ *)'
4.	Cry	': \' (',': '- (', ':' (] ',' (T_T) ',' (; _;) ',' (; _; (; _ :) ',' (; O;) ',' (: _;) ',' (ToT) ',' (T? T) ',': _; ',' ;; ',': n; ',' ;; ' 'Q.Q', 'T.T', 'TnT', 'QQ', 'Q_Q'
5.	Sad	':-(' ':(', ': c', ': c', ': - <', '<', ': - [' ': [', ': - ||' ,

C. SentiWordNet

SentiWordNet a referral from WordNet [9]. WordNet is a dictionary word with different meanings according lexical structure for the purposes of text analysis. SentiWordNet contains a set of words from WordNet labeled sentiment. Type sentiment consists of 3 classes : positive, negative, and neutral. A sentiment score ranging from 0.0 to 1.0. The sum of the three is always worth 1.0.

D. Translation

Due to the limitations of the use of SentiWordNet, we have to translate the data from Indonesian into English. So it can be searched on SentiWordNet sentiment score. We use Google translate to translate it as a media translator. Google translate feels quite good at translating words. It has been able to consider the use of the word in a sentence.

In case of the text in the form of emoticons, emoticons that have been collected in one class also translated according to the emotion class.

E. Pre-processing

1.) Separation

This separation process is known as tokenization. The process started by separating the text into word tokens and symbols. This symbol is generally in the form of punctuation such as periods, commas and question marks. But in this case we try to identify symbols such as emoticons that exist in a sentence. We tried to separate between emoticons and said to then be used in the next process.

2.) POS Tagging

The process of POS (Part of Speech) tagging is the process of marking the word in the text (corpus) in accordance with a particular part of the sentence. POS Tagging aims to determine the exact meaning of a word in a sentence by considering its relationship with the closest words and linked in phrases, sentences, or paragraphs. This process is carried out after the separation of each sentence carried out. There are eight parts of speech roomates are nouns, pronouns, adjectives, verbs, adverbs, prepositions, conjunctions, and interjections.

F. Sentiment Identification

In the previous pre-processing, a sentence will be identified whether the sentence contains emoticons. The results of that process will then proceed to the identification in accordance with certain criteria. Here is the explanation in the form of a diagram. Here is a brief explanation on a Fig. 2.

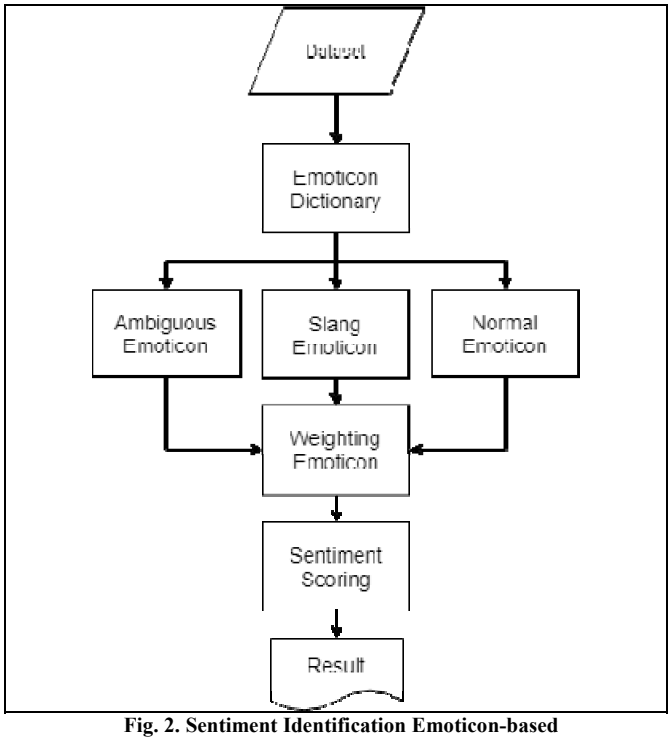


Fig. 2. Sentiment Identification Emoticon-based

1.) Text-based

For text that identified as ordinary word, the process will continue with the search sentiment score of each word in the Sentriwordnet. The end result of this process is the number of positive and negative scores in one sentence. This value is obtained from the sum of positive and negative scores of each word.

2.) Emoticon-based

For text that identified as emoticon, the process is continued in this function. Identified emoticon will be checked. There are three conditions that make a text indicated as emoticon. This condition refers to the emoticon dictionary that was created in the previous method.

a. Ambiguous Emoticon

Ambiguous emoticon is definition of mixed emotions or condition. Condition where the emotion was in the middle between two opposite emotions.

For example emoticon :"), which is a combination of emoticons happy and crying. In this context these emoticons is defined as a condition where the person is experiencing conditions of emotion but happy. This can be found in a tweet that described the touched condition and happy situation at once such as graduation, wedding moments , etc.

TABLE II. AMBIGUOUS EMOTICON SCORE

Ambiguous Emoticon	Emoticons	Positive Score	Negative Score
:')	:)	0.875	0.0
	:'(	0.25	0.0
Result		0.56	0.0

Scores for emoticon of this type is the result of the merger of negative or positive score of both original emoticons. See Table II.

b. Slang emoticon

These emoticons commonly used the statuses on social media. In this case, emoticons written in a format that is not appropriate. As example is emoticon :)))), it is written with ‘)’ excess characters

The addition of ‘)’ could mean conditions 'very'. So that could mean significant emoticons very happy situation. as well as emoticons :(((. That means the condition where a person is in a very sad feeling.

TABLE III. SLANG EMOTICON SCORE

No.	Slang Emoticons	Positive Score	Negative Score
1	:)	0.875	0.0
	:))	0.975	0.0

	:)))...)	1	0.0
2	:(	0.125	0.75
	:((	0.125	0.65
	:(((... (	0.125	0.0

Scores for this emoticon type is addition and subtraction score for each character added. The value of each character added is 0.1. Addition or subtraction depends on the tendency of sentiment on the emoticon. See Table III above.

c. Normal emoticon

Normal emoticons are emoticons that don't have special characteristics and like emoticons in general. Scores for normal emoticon according to the emoticon dictionary that refers to a sentiment score on SentiWordNet.

3.) Weighting Emoticon

In this process, the text of which was identified as an emoticon will going to this process. In this process we perform weighting score in emoticons. We assume that emoticons have more affect on all the tweets and can describe an emotion than ordinary words. So we give more weight that value is double on each emoticon identified.

(1)

$$Positive\ Score = 2 * positive$$

(2)

$$Negative\ Score = 2 * negative$$

4.) Sentiment Scoring

After identifying the text into words and emoticons, the next step is to analyze sentiment score to each sentence m. Sentiment score is obtained from SentiWordNet containing positive and negative scores.

After the positive score of each word in the sentence summed. See Formula. 3. Similarly, the negative score. See Formula. 4 In addition, every word and emoticons that have a value of sentiment also summed. See Formula. 5. These results then that would be a divider in the equation Formula. 6.

(3)

$$S_{positive} = \sum_{i=1}^{n} Positive\ score_i$$

(4)

$$S_{negative} = \sum_{i=1}^{n} Negative\ score_i$$

(5)

$$S_{count} = \sum_{i=1}^{n} Word_i + Emoticon_i$$

The end result of the identification of this sentiment is the number of positive score of equation (3) added of negative scores from equation (4) divided by the number of words in equation (5) in a sentence. For details see the formula in equation. 6

(6)

$$S_{Result} = \frac{S_{positive} + S_{negative}}{S_{count}}$$

G. Sentiment Classification

After going through the process of identification sentiment, we get a score that is the final result to determine the classification of sentiment.

A sentence will have a positive sentiment if the final score is more than or equal to 0.66. The sentence which has the final score less than or equal to 0:34 will be categorized as negative sentiment. And the last, if final score is more than 0:34 and less than 0.66 will be categorized as neutral sentence.

(5)

$$S_{Sentences} \begin{cases} Positive\ ,if\ S_{Result} \geq 0.66 \\ Negative\ ,if\ S_{Result} \leq 0.34 \\ Neutral\ ,if\ S_{Result} > 0.34\ and\ S_{Result} < 0.66 \end{cases}$$

IV. RESULT

TABLE IV. SENTIMENT CLASSIFICATION RESULT

	Positive	**Neutral**	**Negative**
Accuracy	0.77	0.60	0.78
Average		0.74	

Table IV shows the results of the experiment for each type of dataset are positive, negative, and neutral. This result shows the accuracy for each type of sentiment. The average of the three types of 0.74.

The table below shows comparison results of accuracy, specify, precision and recall of algorithms that use emoticons-based and without using emoticons-based.

TABLE V. COMPARISON RESULT

	SentiWordNet	Emoticon SentiWordNet-based
Accuracy	0.59	0.74
Specify	0.58	0.73
Precision	0.59	0.72
Recall	0.59	0.74

In Table V it can be seen that the accuracy of the algorithm that uses emoticons SentiWordNet-based reach 0.74. This value is much larger than than the usual accuracy of the algorithm SentiWordNet which was only 0.59.

As well as on the value Specify, precision and recall. Third value derived from the results of SentiWordNet algorithm is much greater than the value of the ordinary SentiWordNet algorithm.

V. CONCLUSION AND FUTURE WORK

In this paper, we try to identify the sentiment of a sentence by considering emoticons. We assume that an emoticon expresses more obvious emotions than words. So we give more weight to the emoticons. From the experiments that have been done, we get an accuracy of 0.74. Overall these results are much better than using SentiWordnet algorithm withouut emoticons-based, which only has an accuracy of 0.59.

For the study in future work is the use of WSD (Word Sense Disambiguation) [13] [14] in a word in order to know the exact meaning in a word so the sentiment score will give appropriate results.

REFERENCES

[1] M. Taboada, J. Brooke and M. Tofiloski, "Lexicon-based methods for sentiment analysis," Computational linguistics, pp. 267-307, 2011.

[2] Ohana B. and B. Tierney, "Sentiment classifiation of reviews using SentiWordNet," in 9th. IT & T Conference 2009.

[3] E.W. Pamungkas, R. Sarno, A. Munif, "B-BabelNet: Business-Specific Lexical Database for Improving Semantic Analysis of Business Process Models, Telecommunication, Computing, Electronics and Control (TELKOMNIKA)", Vol 15, No 1, 2017. DOI: http://dx.doi.org/10.12928/telkomnika.v15i1.3176

[4] S. Baccianella, A. Esuli and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining." LREC, vol. 10, pp. 2200-2204, 2010.

[5] W. Devid Haryalesmana, N. Azhari S, "peringkasan Esktraktif sentiment on Twitter Using TF-IDFdan Hybrid Cosine Similarity", Indonesian Journal of Computing and Cybernetics Systems, Vol.10, No.2, July 2016, pp. 207 ~ 218

[6] M. Edison and A. Aloysius "based Lexicon Acronyms and Emoticons of Sentiment Classification Analysis (SA) on Big Data", International Journal of Database Theory and Application, Vol.10, No.7 (2017), pp.41-54 (2017).

[7] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[8] M. Young, The Technical Writer's Handbook. Mill Valley, CA: University of Science, 1989.

[9] WordNet, A Lexical Database for English, "https://wordnet.princeton.edu/" [Accessed 28 April 2018]

[10] List of emoticons, "https://en.wikipedia.org/wiki/List_of_emoticons/" [Accessed 28 April 2018]

[11] F. H. Rachman, R. Sarno, C. Fatichah, "Music Emotion Classification based on Lyrics-Audio using Corpus-Based Emotion", International Journal of Electrical and Computer Engineering (IJECE), Vol 8, No 3, pp. 1720-1730, Juni 2018. DOI: 10.11591/ijece.v8i3.ppi-720-1730.

[12] F. H. Rachman, R. Sarno, C. Fatichah, "CBE : Corpus-Based off Emotion for Emotion Detection in Text Document". The 3rd International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE), Pages: 331 – 335, 2016. DOI: 10.1109/ICITACEE.2016.7892466.

[13] B. S. Rintyarna., R. Sarno, "Adapted Weighted Graph for Word Sense disambiguation", The 4th International Conference on Information and Communication Technology (ICoICT), 2016. DOI: 10.1109 / ICoICT.2016.7571884.

[14] B. S. Rintyarna, R. Sarno, C. Fatichah, "Enhancing The Performance Of Sentiment Analysis Task On Product Reviews By Handling Both Local And Global Context". International Journal of Information and Decision Sciences. 2018 Vol. 11.

Word Sense Disambiguation in Bahasa Indonesia Using SVM

Edi Faisal
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
faisal@dinus.ac.id

Farza Nurifan
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
farzanurifan@gmail.com

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

***Abstract*—Currently, the use of Indonesian language on the internet is growing rapidly and many of the existing sentences contain ambiguous words. In Natural Language Processing the problem to find out the meaning of an ambiguous word is called Word Sense Disambiguation. Word sense disambiguation is a problem about how we know the meaning of an ambiguous word in a given sentence. Many uses if we can solve word sense disambiguation problems such as can be used for text classification, text clustering and for machine translation. In this paper, we propose the use of SVM algorithm with TF-IDF as the feature extraction method and Wikipedia as the training data to solve the WSD problem of Indonesian language. The results of our proposed method reach an accuracy level of 0.877.**

Keywords— Word Sense Disambiguation; SVM; TF-IDF; Wikipedia

I. INTRODUCTION

Internet usage in Indonesia is currently growing very fast. Lots of information in the form of sentences spread through social media by the Indonesian people. From that sentence, we can analyze various things one of them is the analysis of sentiment[1]. But if there is an ambiguous word in a sentence then we will be difficult to analyze it. The field that handles the problem is word sense disambiguation (WSD) where we make the computer understand what the ambiguous word really means from a sentence. For example, in the sentence:

1. "Malam ini bulan purnama terlihat indah."
2. "Malam hari di bulan Juni terlihat indah."

From the above sentence, the word "bulan" in Bahasa Indonesia has multiple meaning. In the first sentence "bulan" means "Earth satellites" and in the second sentence "bulan" means "Time".

Implementation of word sense ambiguation is widely used in information retrieval, lexicography, knowledge mining / acquisition, and semantic interpretation, and in current emerging studies such as bioinformatics and the Semantic Web. Word sense disambiguation also becomes the main solution in overcoming the ambiguity of words or sentences such as text classification [2] and text clustering [3].

Many researchers have proposed various methods to solve word sense disambiguation in various languages outside English such as building a corpus for supervised word sense disambiguation for the Arabic language [4] with Arabic wordnet mapping method and Wikipedia to select Wikipedia article that matches with wordnet. Furthermore, the cross-lingual method is used to measure the similarity between the feature of Wikipedia article with wordnet. In addition, WSD is also used for classification in Turkish [5]. Classification is done with 6 (six) machine learning method that is C4.5, Random forest, Rocchio, Naïve Bayes, KNN, linear and multilayer perceptron. WSD is also used to build corpus word sense disambiguation in Hindi [6], [7]. Sense annotated Hindi corpus developed from Hindi wordnet which has been available in porta Technology Development for Indian Languages (TDIL). Word sense disambiguation for Assam languages [8] using machine learning Naive Bayes Classifier. While Rini Wongso [9] using machine learning for news classification in the Indonesian language regardless of word ambiguation. The TF-IDF and SVD algorithms are used for feature selection and classification methods using Multinomial Naïve Bayes, Multivariate Bernoulli Naïve Bayes, and Support Vector Machine. Benchmarking of this method is used to find the best classification method.

Navigli [10] makes 3 main approaches in word sense disambiguation: knowledge-based, unsupervised and supervised approaches. According to Navigli [10] supervised approached is the best approach among the three approaches above, provided that the training data used is data with good quality. With reference to this study using data training from the Wikipedia site with the consideration he is a repository of the largest knowledge encyclopedias available for free in 282 languages, has high editorial quality [11], has many uses like to improve semantic analysis [12], and already has a category that can be used as a label. There are many another approach for word sense disambiguation. For example, using adapted weighted graph [13] or using degree and dependency structure of association between sentences [14] for word sense disambiguation.

In this paper, we propose the use of SVM algorithm with TF-IDF as the feature extraction method and Wikipedia as the training data to solve the WSD problem of Indonesian language. We use SVM because in our test it has been shown that this is the best classification that has the highest accuracy. First, we search Wikipedia articles to use it as the training data. The Wikipedia articles selection is based on the category label

from the articles itself. The articles are then preprocessed to make the words variation to be fewer. After that we use TF-IDF to convert the articles into classifiable features and then we determine the meaning of an ambiguous word in a sentence with Support Vector Machines classifier.

This paper is structured as follow; Section II is the Methodology; Section III is the Result; Section IV is the Conclusion

II. METHODOLOGY

In this paper, we divide our method process into five parts; first is Wikipedia articles selection; second is training and testing data; third is Preprocessing; fourth is Machine Learning using SVM; fifth is performance measure.

A. Wikipedia articles selection

The selection for Wikipedia article is done by first searching for a predefined ambiguous word, in this case, "motor" which has the ambiguous meaning of motorcycle (vehicle) and propulsion (electric). It is also the word "bulan" that has ambiguous meanings of time and earth satellites.

Next look for Wikipedia articles that match the ambiguous words above by looking at the categories in the Wikipedia article which will be used as a label of the sentence that contains the word ambiguous. From the articles obtained, then taken a sentence that contains the word ambiguous and grouped into two files according to the category. From the file, the sentences that obtained from Wikipedia are then given a label to distinguish the meaning according to the group of ambiguous words. The flowchart of Wikipedia article selection can be seen in Figure 1.

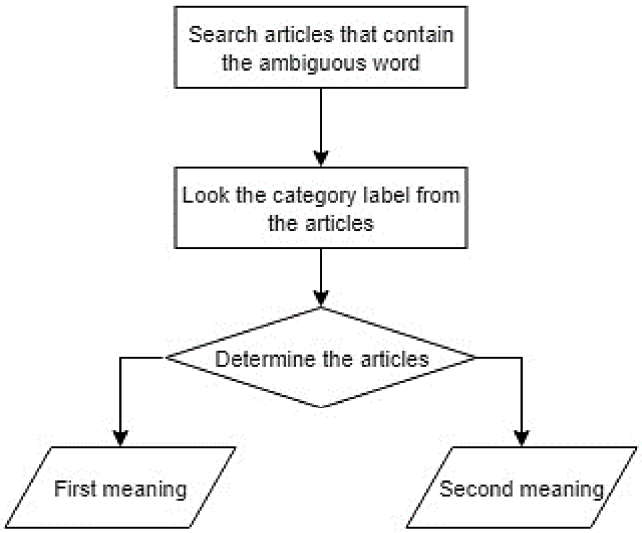


Figure 1. Wikipedia article selection

B. Training and Testing data

From the search results on the Wikipedia site obtained a sentence that contains the ambiguous word and has been labeled. This label is used to distinguish the meaning of the ambiguous word. Having obtained a sentence that contains the word ambiguous the next list of sentences is used as data training that will be used in machine learning. The use of Wikipedia as training data because Wikipedia is known as a site that has high editorial quality [11] and has a category that can be used as a label.

While as testing data searched from the article some internet sites such as from news sites and special sites to discuss articles that contain the above ambiguous words. From the site of the site, as well as on the training, then sentence sentences containing the word ambiguous "motor" and "bulan". The sentences then labeled as in the training data. Labeling on the data testing is done manually. This label will not be processed in machine learning but will be used as a comparison of machine learning process prediction results.

C. Preprocessing

The training data that obtained from Wikipedia is then preprocessed to eliminate the noise word of the sentence which has no significant meaning in the sentence [15]. The preprocessing stage is:

1) Lowercase: This process is done by changing all words into lowercase. This is so that each word is unique without any distinction between uppercase and lowercase.

2) Remove punctuation: Because punctuation is considered to have insignificant meaning in the sentence, so it needs to be eliminated.

3) Stemming: This process is done by changing the existing word into a basic word, especially for the verb.

4) Remove stop words: This process is done by eliminating words that have no significant meaning in sentences such as words that, and, others and the like.

Training data is done by preprocessing process with 4 stages above because the data is taken from good sources in language review.

For the testing data is still done additional stage preprocessing, because the search results of data from multiple sites allows there are data that have words that are not standard or contain abbreviations. Here's an additional process of preprocessing in data testing.

5) Normalized slang word: Slang word is a non-standard word used by people in unofficial communication. These words need to be matched with the standard word.

6) Normalized abbreviation word: Normalization of words containing abbreviations is only done on words that have meaning and influence in the sentence. While words that have no meaning and influence in the sentence will be deleted.

The example of preprocessing result can be seen in Table 1.

TABLE 1. PREPROCESSING RESULT

Input	Output
saya pergi ke bulan	pergi bulan
saya menaiki sepeda motor	naik sepeda motor

D. Machine Learning using SVM

Data that has been preprocessed then will be processed in machine learning with stages as below

1) TF-IDF Feature Extraction: Term Frequency-Inverse Document Frequency is a feature selection that will change the word in the document into a value (weight) in a statistic where the weight indicates how important the word is in a document [16]. The Stages performed in the weighting by using the TF-IDF are as follows:

a) Tokenization: Preprocessing training data is transformed into uni-gram and bi-gram. Data used as uni-gram is meant to know the weight of each word. While made bi-gram to know the weight when two words have a related meaning.

Examples of uni-gram and bi-gram tokenization:
Sentence: bertemu bulan januari
Uni-gram: {bertemu, bulan, januari}
Bi-gram: {bertemu bulan, bulan januari}
Merge: {bertemu, bulan, januari, bertemu bulan, bulan januari}

The flowchart of tokenization process shown in Figure 2.

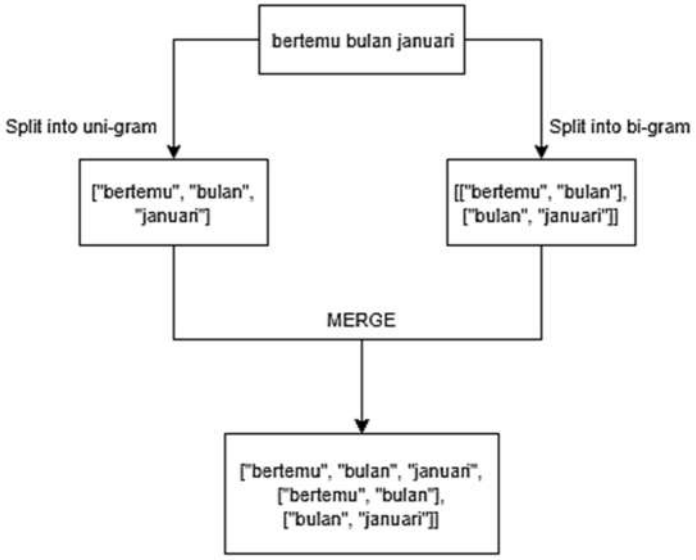


Figure 2. Tokenization

b) Term Frequency: Term frequency is done by counting the number of words that appear in a document [12]. The word that often appears indicates that the word is important. Term Frequency is calculated using the formula:

$$TF_{(t,d)} = \frac{n_{t,d}}{n_d} \qquad (1)$$

Formula (1), $TF_{(t,d)}$ is the number of terms t in the document $n_{t,d}$ is the number of occurrences of term t in document d. Whereas n_d is the sum of the terms in the document d

c) Inverse Document Frequency: Inverse document frequency is a measure that determines how important a word is in all documents. It shows often or rarely a word appears in all documents. From this it is seen that preprocessing becomes important, as there are several words that very often appear but have no meaningful meaning like and, which, as well as, etc.

For the calculation of IDF in this paper will be used the formula:

$$IDF(t,d) = \log\left[\frac{(1+n)}{(1+df(t,d))}\right] + 1 \qquad 2)$$

From the above formula (2) it is known that *IDF (t,d)* is an inverse document frequency of term t in document d. *n* is the number of documents used, whereas $df(t,d)$ is the number of documents that contain term t. The addition of 1 of the formula (2) is intended for the sum to not be 0 to avoid the divisor 0.

While constant 1 is meant that the result of idf is at least 1.

After calculating the value of TF and IDF then calculated the value of TF-IDF with the following formula:

$$TF - IDF_{(t,d)} = TF_{(t,d)} . IDF(t,d) \qquad 3)$$

d) Normalize TF-IDF: Normalization is done using the L2 norm so that the weight of tf-idf for each term has a scale weight of 0-1, see the following formula

$$v_{norm} = \frac{v}{\|v\|^2} = \frac{v}{\sqrt{{v_1}^2 + {v_2}^2 + \cdots + {v_n}^2}} \qquad 4)$$

The example of TF-IDF weighting can be seen in Table 2.

TABLE 2. TF-IDF WEIGHTING

No	Term	TF				TF-IDF	
		K1	K2	DF	IDF	K1	K2
1.	temu	1	0	1	1.17609	1.176090	0
2.	bulan	1	1	2	1	1	1
3.	januari	1	1	2	1	1	1
4.	temu bulan	1	0	1	1.17609	1.176090	0
5.	bulan januari	1	1	2	1	1	1
6.	berangkat	0	1	1	1.17609	0	1.17609
7.	berangkat bulan	0	1	1	1.17609	0	1.17609

2) Suppor Vector Machine (SVM): Support vector machine is a machine learning model used to predict classification analysis and regression analysis [17]. To

perform the classification, SVM will search for the best hyperplane function as a separator between two data classes.

In this research, we will use linear kernel function of SVM with the consideration that the data to be classified can be separated by a hyperplane.

E. Performance Measure

The performance measure is determined by 4 (four) formulas: precision, recall, F1Score and accuracy.

Precision is a measure of the proximity of the correct ambiguous sentence prediction to the overall outcome of the correct ambiguous sentence. Precision can be written in the following mathematical sentences

$$Precision = \frac{Tx}{Tx + Fx} \quad (6)$$

Recall is the correct number of ambiguous sentence predictions versus the total number of ambiguous sentences in document x. Recall can be written in the following mathematical sentences

$$Recall = \frac{Tx}{Tx + Fy} \quad (7)$$

F1Score is a measure of test accuracy by considering the precision and recall results. F1Score can be written in the following mathematical sentence

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (8)$$

The accuracy of the test is the ability to distinguish the correct ambiguous sentence predictions from the whole sentence tested. Accuracy can be written in the following formula:

$$Accuracy = \frac{Tx + Ty}{Tx + Fx + Ty + Fy} \quad (5)$$

where
Tx = the amount of true prediction of the first sense
Fx = the amount of false prediction of the first sense
Ty = the amount of true prediction of the second sense
Fy = the amount of false prediction of the second sense

For example, the word "bulan" has two senses, the first sense is "time" and the second sense are "earth satellite". We calculate for each sense. When the first sense is "time" then the second sense is "earth satellite" and vice versa.

III. RESULT

In this paper, we use python as a programming language to implement the method we propose. For preprocessing Indonesian documents, we use Pysastrawi python library and for TF-IDF and SVM we use scikit-learn python library. Pysastrawi is a python library that allows us to reduce inflected words in Bahasa Indonesia to their base form, for example, the word "menahan" will be changed to "tahan". We use two ambiguous words as the testing data; first is "motor" with sense vehicle and propulsion (electric); second is "bulan" with sense time and earth satellite.

The confusion matrix result for "motor" can be seen in Table 3 and for "bulan" in Table 4. The performance measure of this experiment can be seen in Table 6.

TABLE 3. CONFUSION MATRIX OF "MOTOR"

Prediction	Actual	
	Vehicle	Propulsion
Vehicle	63	11
Propulsion	2	51
Total Data	65	62

TABLE 4. CONFUSION MATRIX OF "BULAN"

Prediction	Actual	
	Time	Earth Satellite
Time	46	7
Earth Satellite	8	43
Total Data	54	50

TABLE 5. EXPERIMENT RESULT

Result	Bulan		Motor		Average
	Time	Satellite	Vehicle	Propulsion	
Precision	0.852	0.860	0.969	0.823	**0.876**
Recall	0.868	0.843	0.851	0.962	**0.881**
F1 score	0.860	0.851	0.906	0.887	**0.876**
Accuracy	0.897	0.897	0.855	0.855	**0.877**

IV. CONCLUSION

The methodology that we propose to solve word sense disambiguation in Bahasa Indonesia using Wikipedia and Support Vector Machine reach an accuracy rate of 0.877. This accuracy is good considering that there is no other method proposed for word sense disambiguation in Bahasa Indonesia. For further research, the senses of ambiguous words can be added so that not only have two senses.

REFERENCES

[1] B. S. Rintyarna, R. Sarno, and C. Fatichah, "Enhancing the performance of sentiment analysis task on product reviews by handling both local and global context," *Int. J. Inf. Decis. Sci.*, vol. 11, 2018.

[2] E. Gabrilovich and S. Markovitch, *Journal of artificial intelligence research JAIR : an international electronic and print journal.*, vol. 34, no. 1. AI Access Found, 2009.

[3] A. Saif, M. J. Ab Aziz, and N. Omar, "Reducing explicit semantic representation vectors using Latent Dirichlet Allocation," *Knowledge-Based Syst.*, vol. 100, pp. 145–159, May 2016.

[4] A. Saif, N. Omar, U. Z. Zainodin, and M. J. Ab Aziz, "Building Sense Tagged Corpus Using Wikipedia for Supervised Word Sense Disambiguation," *Procedia Comput. Sci.*, vol. 123, pp. 403–412, Jan. 2018.

[5] O. Acikgoz, A. T. Gurkan, B. Ertopcu, O. Topsakal, B. Ozenc, A. B. Kanburoglu, I. Cam, B. Avar, G. Ercan, and O. T. Yildiz, "All-words word sense disambiguation for Turkish," in *2017 International Conference on Computer Science and Engineering (UBMK)*, 2017, pp. 490–495.

[6] S. Singh and T. J. Siddiqui, "Sense Annotated Hindi Corpus," in *2016 International Conference on Asian Language Processing (IALP)*, 2016, pp. 22–25.

[7] D. Wali and N. Modhe, "Word Sense Disambiguation Algorithms in Hindi," 2015.

[8] J. Sarmah and S. K. Sarma, "Word Sense Disambiguation for Assamese," in *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, 2016, pp. 146–151.

[9] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, and Rudy, "News Article Text Classification in Indonesian Language," *Procedia Comput. Sci.*, vol. 116, no. C, pp. 137–143, 2017.

[10] R. Navigli and Roberto, "Word sense disambiguation," *ACM Comput. Surv.*, vol. 41, no. 2, pp. 1–69, Feb. 2009.

[11] S. Tonelli, C. Giuliano, and K. Tymoshenko, "Wikipedia-based WSD for multilingual frame annotation," *Artif. Intell.*, vol. 194, pp. 203–221, Jan. 2013.

[12] E. W. Pamungkas, R. Sarno, and A. Munif, "B-BabelNet: Business-Specific Lexical Database for Improving Semantic Analysis of Business Process Models," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 15, no. 1, p. 407, Mar. 2017.

[13] B. S. Rintyarna and R. Sarno, "Adapted weighted graph for Word Sense Disambiguation," in *2016 4th International Conference on Information and Communication Technology (ICoICT)*, 2016, pp. 1–5.

[14] K. Mishina, S. Tsuchiya, and H. Watabe, "Word sense disambiguation of adjectives using dependency structure and degree of association between sentences," in *2017 International Conference on Asian Language Processing (IALP)*, 2017, pp. 342–345.

[15] S. Vijayarani, M. R. Janani, and A. Professor, "TEXT MINING: OPEN SOURCE TOKENIZATION TOOLS – AN ANALYSIS," *Adv. Comput. Intell. An Int. J.*, vol. 3, no. 1, pp. 37–47, 2016.

[16] H. C. Wu, R. W. P. Luk, K. F. Wong, and K. L. Kwok, "Interpreting TF-IDF term weights as making relevance decisions," *ACM Trans. Inf. Syst.*, vol. 26, no. 3, pp. 1–37, Jun. 2008.

[17] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A Practical Guide to Support Vector Classification."

Developing Word Sense Disambiguation Corpuses using Word2vec and Wu Palmer for Disambiguation

Fadli Husein Wattiheluw
Department of Informatics
Institut Teknologi Sepuluh Nopember, Surabaya , Indonesia
fadliwattiheluw1994@gmail.com

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember, Surabaya , Indonesia
riyanarto@if.its.ac.id

Abstract*— In computational linguistics, meaning disambiguation is an open problem of natural language processing in the form of the process of identifying the meaning of the word polysemy used in a sentence. Resolving this problem, among others, has an impact on search engine relevance, anaphoric solving, coherence or cohesion, and inference or conclusion. Therefore, a study is needed that studies to find the meaning of a correct word on a topic. So that it affects the topics discussed in a sentence to find the true meaning. In this study, we focused on finding the meaning of words in a corpus-based sentence using word2vec and wu palmer. The word2vec algorithm is used to construct word vectors contained in sentences and wu palmer as an addition to new words that are not contained in the corpus, by assessing hypernym, meronym, and hyponym between words in sentences. The experimental results show that by adding a new word using wu palmer on corpus it can increase the precision value of 0.8232 in an introduction to a sentence contained in a topic, compared to not using the addition of a new word.

Keywords—Word sense disambiguation, hyponym, meronym, hypernym, wu palmer, word2vec.

I. INTRODUCTION

Also known as sentiment analysis is opinion mining refers to process determining opinions or emotions expressed in a text about the subject. Although sentiment analysis is a field of research recently, which was introduced in 2001 [1] This raised a lot of interest and many applications to know the sentiments on the opinion of users, for example, in the product reviews, news, twitter, and blog [2].

In determining better sentiments in user comments, the word disambiguation plays a role in sentiment to determine the meaning of a word in a sentence. Word Sense Word Sense Disambiguation (WSD) is an ability to identify the meaning of words in computing [3]. There are many polysemic words that have different meanings for each topic. Sometimes it is not easy for a computer to identify the meaning of some polysemic words on a particular domain so that it requires a model that is used to overcome the problem of polysemic words.

For example, there were two sentences:

(1) I and my friend stayed in room 201 for 2 nights.

(2) My friend was treated in room 201 for 2 nights.

In the sentence, there is the word “room” which has a different meaning in hotel and hospital domain. The word “room” in the first sentence describes room for traveling, while. The word ' room ' in the second sentence explains the care of the sick. Computers will have difficulty determining the meaning of a sentence well if there is a polysemic word. Therefore, it has become a major problem in several studies of natural language processing. Techniques of word sense disambiguation is an automatic way to determine meaning word a context in opinion. Generally, WSD is identifying which sense of a word is used in a sentence when the word has multiple meanings.

The last few years there is a research about WSD, such as that done by Bagus Setya [4]. This research is concerned with the improvement of model graph-based approach WSD, weight graph extracted by using some measure of similarity (i.e.: Leacock & Chodorow, Wu Palmer, Resnik, Lin And Jiang & Conrath). to improve the model by increasing the lesk algorithm with adapted lesk [5]. As for other studies conducted by Amita, Devendra, and Sonakshi [6] distinguish words in wordnet uses fuzzy semantic relations. Where the early stages of defining words in the lexical categories. Furthermore, Identify the words that stand out in the context of the collective. After narrowing the categories, understanding of the word is found using a modified lesk algorithm. As for other studies conducted by Su and Thanda [7] focused on both supervised and knowledge-based approaches. With the new coefficient-based WSD algorithm proposed to overcome match vocabulary problems. External knowledge resources corpus and wordnet are used as a repository of sense by linking with the new WSD algorithm to consider additional semantics for WSD.

Other studies related to the development of corpus conducted by Fika, Riyanarto, and Chastine [8]. They built the corpus to detect emotions in documents using a model called Corpus-Based of Emotion (CBE). CBE developed from Affective Norms for English Words (ANEW) and Wordnet Affect Emotion (WNA) with the term similarity and distance approaches the size of the node. in addition to the research conducted by Fika, automatically adding new words that are not found on the CBE corpus using Latent Dirichlet Allocation (LDA). As for other studies to classify emotions in the music domain using lyrics and audio as a feature [9]. The lyric feature is extracted from text data and audio features extracted from audio signal data. In the classification of emotions, the emotional corpus is needed for the extraction of lyrical features. Corpus-Based Emotion (CBE) managed to increase the value of F-Measure for classification emotions in text documents. Music documents have an unstructured format compared to article text documents. So that it requires a good preprocessing and conversion process before the classification process. The best test results for the classification of musical emotions are the application of the Random Forest method for lyrics and audio features.

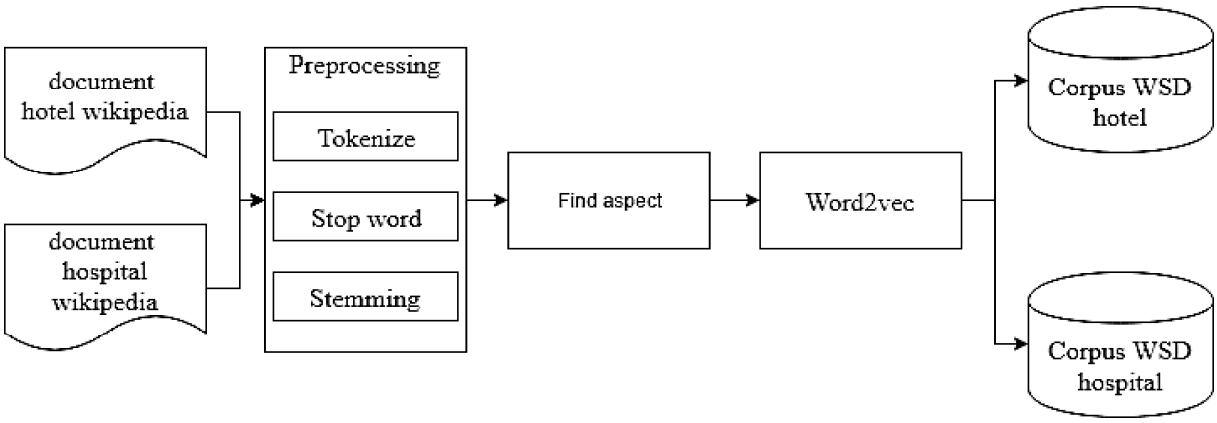


Fig. 1. Step develop corpus WSD

Research conducted by Endang [10]. They proposed a new lexical database called B-BabelNet. The model proposed to improve analysis of semantic business process model. They try to map Wikipedia pages to WordNet database but only focus on words related to the business domain. In addition, to enrich vocabulary in the business domain. They also use terms in specific online business dictionaries. The results in the disambiguation process using B-BabelNet show an increase in the accuracy of meaning disambiguation, especially in certain matters related to business and industrial domains.

Problems often encountered in natural language processing are in determining the meaning of a polysemic word which means it will be different in a topic or domain. So it requires a sense of disambiguation to distinguish corpus-based words for each domain. Therefore, this study proposes that the WSD corpus is built using the Word2vec algorithm for the initial stage of the WSD corpus formation and adds a new word using Wu Palmer. To expand the WSD corpus built for a particular domain. There are several stages to build a WSD corpus for a particular domain. First, every training document taken from Wikipedia will be preprocessed. Then each word will be carried out by the training process using the word2vec algorithm to build vector terms. To be able to find out the topic on the testing document, the similarity value will be calculated based on the word vector on each topic. to compare documents with certain topics using several important words as aspects that explain the topic. To find the important word, we use the term frequency technique to find important words with the frequency of occurrences on the topic. The words that are not in the WSD corpus are built on certain domains. To overcome words that are not in the corpus, we use the Wu Palmer algorithm to find the similarity of the meaning of the word. To add new words to the WSD corpus by calculating the value of the similarity between words based on hypernym, hyponym, and meronym values.

This paper is organized as follows: Part II describes the stages of creating a corpus and determines the meaning of words on the topic. Section III describes the analysis and evaluation of the results of the proposed method. And the last part IV describes the conclusions from the results of the experiment.

II. METHODS

In developing the WSD corpus, the method proposed is divided into two main parts: (A) At the stage to build a WSD corpus that uses Wikipedia as training data. Training data will be preprocessed, search for important words and build vectors for each word on the topic using word2vec. (B) At this stage, we automatically add new words that are not in the WSD corpus. The process of adding new words using the Wu Palmer algorithm by considering the value of similarity between hypernym, hyponym, and meronym. In this study will focus on hotel and hospital domains in overcoming WSD.

A. Develop corpus WSD

The training document will be carried out in the preprocessing process (ie tokenize, stemming, and stop word) to eliminate unnecessary words in building WSD corpus. After the preprocessing process, the next is to determine important words based on the number of events on the topic of hotels and hospitals using the term frequency technique. After determining important words for each topic, we build vectors for each word using the word2vec algorithm. Word vectors contained in the WSD corpus are used to distinguish the same word for different topics.

Figure 1 illustrates the initial development stage for the WSD corpus in handling polycemic words. Where training data is taken manually from Wikipedia with hotel and hospital keywords. The number of datasets taken about hotels and hospitals with a total of 134 documents used as training data to develop the WSD corpus. After collecting data for training, the next stage will be carried out. Data processing starts with tokenizing, stemming, and deleting irrelevant words in the WSD corpus. The next stage, the preprocessing results document will be carried out in the process of finding important words in each document that is used as an aspect of the topic of hotel and hospital. After finding the word as an aspect, then from the training data will be carried out in the process of vector formation using the skip-gram model for each word in hotel and hospital documents. The last stage, every word that has been trained will have vector values for hospital documents and hotels

that will be stored as word sense. As for the explanation of each process:

1) Preprocessing

Training document data from Wikipedia will be broken down into a term called tokenize. Furthermore, unimportant terms contained in the tokenize results to be deleted are called stop words such as a the, in, for, etc. Data has been done the stop word process then undertaken the process of stemming, whereby every word will be returned to basic Word.

For example, there is a document "A hotel is an establishment that provides paid lodging on a short-term basis". Document conduct done process tokenize, where a sentence would be broken into words. Tokenize results "A", "hotel", "is", "an", "establishment", "that", "provides", "paid", "lodging", "on", "a", "short", "term", "base". After done tokenize on a document. The next step, namely the stop word process to eliminate unimportant words such as on, a, is, that. The results of the stop word as follows: "hotel", "establishment", "provides", "paid", "lodging", "short", "term", "base". And the last stage of preprocessing that is stemming. Whereby, every word quickly became basic words into: "hotel", "establish", "provide", "paid", “lodge”, “short”, "term", "stale". This preprocessing step we use natural language toolkit python library in helping our work.

2) Find word aspect in the topic

After preprocessing data on hotel and hospital training from Wikipedia. The next step, we look for candidate words that are used as aspects presented about hotels and hospitals. In defining word used as an aspect to present hotel or hospital. we use the Term Frequency (TF) to find candidate words that are used as aspects. Where words have more frequent occurrences for a topic in hotel and hospital. To search for important words about hotel and hospital topics, we use the following equation:

$$tf(w_t, d) = \frac{\sum_{t=1}^{n} w_{t,t}}{N} \quad (1)$$

Where $tf(w_t, d)$ is the number of frequencies of words that appear on a document, N is a number of words in document d and $w_{t,t}$ is the number of occurrences of words w_t in document d.

TABEL 1. TERM FREQUENCY

Topic	Term			
	establish	provide	paid	Room
Hotel	1	3	4	6

On table 1 displays term frequently appears on the topic hotel. Where the term "space" has a number of occurrence frequencies greater than other words in the hotel topic. The value of "room" term frequency using equation (1) has a value of 0.428 which is used as an aspect presented about the hotel. The results of searching for words as aspects that we use on the topic of hospitals and hotels can be seen in table 2.

3) Word2vec

After doing the stages to look for word candidates as aspects that present topic of hotel or hospital. next is to create a vector for each word that presents words about the topic of hotels and hospitals. This stage we use script-gram model algorithms on word2vec [11] to find the vector representation of word *w(t)* to predict surrounding words in a sentence. As shown in figure 2.

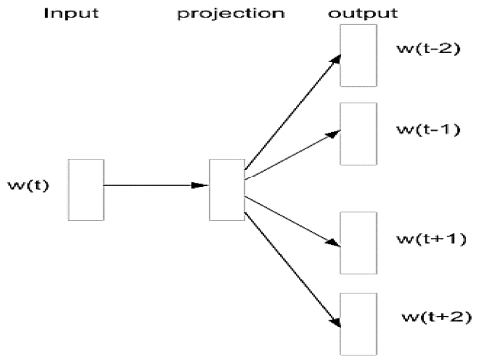


Fig. 2. The Skip-gram model architecture.

To find a word representation is useful for predicting the surrounding words in a sentence or document. given a sequence of word training $w_1, w_2, w_3, \ldots, w_t$. In maximizing the average log probability using the equation [11] :

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0} \mathrm{Log}\, P(w_{t+j}|w_t) \quad (2)$$

Where c the context is the measurement of training (that could be a function of the word center w_t). Formulation of Skip-gram basis defines $P(w_{t+j}|w_t)$ using softmax function:

$$p(wO|wI) = \frac{\exp(v'_{wo}\, v^T_{wI})}{\sum_{w=1}^{W}\exp(v'_w\, v^T_{wI})} \quad (3)$$

where v_w and v'_w are “input” and “output” vector representations of w. w is a number of words in the vocabulary. The results of vector formation for each word for hotel and hospital topics, we use the gram-skip model on word2vec as shown in table 2, where every word has a vector representation. Utilizing the Word2vec algorithm, we use the Gensim library in python to help our work build vectors for each word on the topic of hotel and hospital.

4) Polysemic word sense

After creating a vector for each word presented on the topic of hotels and hospitals using word2vec. Then each word will be saved into the bin file that is used to load a collection of word2vec result vector words. With a vector which is used as a corpus word sense disambiguation to the topic of hotel and hospital. As for example words and vector are stored on each topic of hotel and hospital that can be shown in table 2.

TABEL 2. VECTOR WORD ASPECT REPRESENTATION HOTEL AND HOSPITAL

Topic	Aspect Term	Vector Word
Hospital	Room	[0.2520761 -0.08401345 -0.12540267 ... 0.10789914 0.0204985 0.36947712]
	laboratory	[0.27022526 -0.10543724 -0.13813923 ... 0.0681444 -0.01654486 0.42020062]
	Service	[0.46512705 -0.12488033 -0.18997538 ... 0.13502474 0.02251283 0.5396639]
	Sleep	[0.49783045 -0.20110206 -0.29111138 ... 0.15726066 0.0426247 0.65003633]
Hotel	Room	[-0.01425708 0.01081498 0.02258818 ... 0.02870315 0.00616242 0.00703874]
	Area	[0.02792124 0.00325425 0.02134224 ... 0.03123865 0.00904845 0.00219979]
	Service	[0.0408654 0.02672396 0.03365852 ... 0.01243731 0.00323341 -0.02636981]
	Facility	[0.0261229 0.01244247 0.01922303 ... -0.02984453 -0.02988675 0.03401406]

Each word has a different vector for hospital and hotel topics. Every word has a different vector w_t although the same word on different topics. For example, the word "room" has a vector [0.2520761-0.08401345-0.12540267 ... 0.10789914 0.0204985 0.36947712] in the hospital. But the vector of the word "room" in the hotel domain will be different from the vector of the word "room" in the hospital domain. To find out the topic in the sentence, we use the following equation:

$$P(d|z) = \frac{\sum_{t=1}^{n} SIM_{(wt,z)}}{N} \quad (4)$$

Where N is the number of words in the document, $P(d|z)$ is the probabilistic document on the topic, dan $SIM_{(wt,z)}$ is the value of the similarity of the word w_t with the topic.

B. Automatic expand polysemic WSD

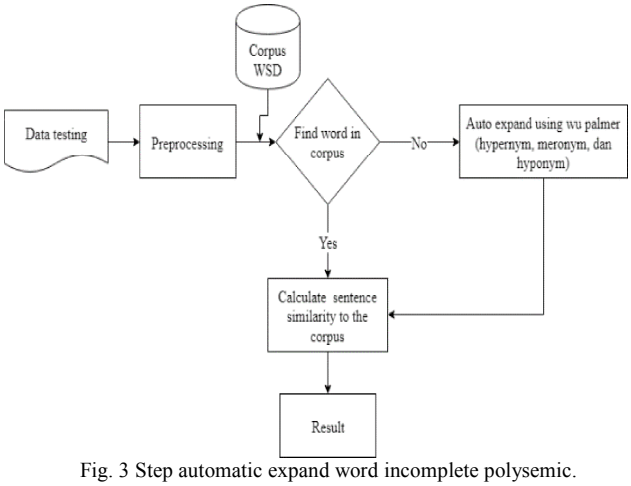


Fig. 3 Step automatic expand word incomplete polysemic.

5) Auto expand corpus

At this stage, we process the addition of new words if the search term is not found in the WSD corpus. Where a new word will do the search process the biggest similarity values using the algorithm Wu Palmer [12] with the equation:

$$WUPw_t = \frac{2 \times N}{N1 + N2} \quad (5)$$

Where N is the parent of the first word of a line $N1$ with the second word $N2$, $N1$ is the number of the line into the first word, dan $N2$ is the number of the line into the second word.

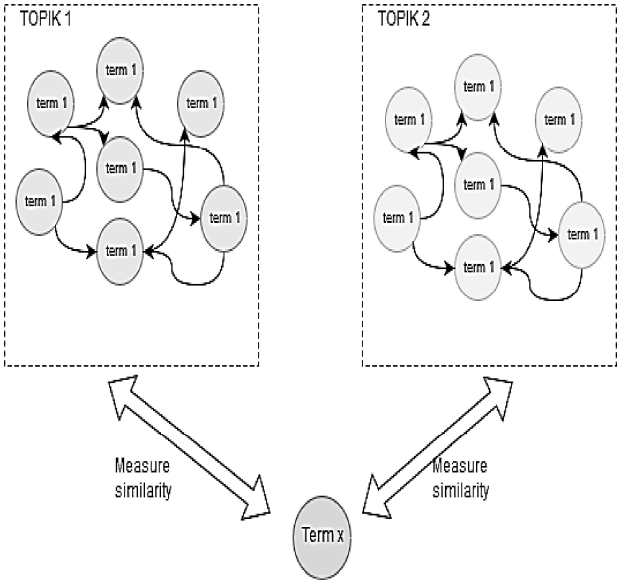


Fig. 4 Examples of adding new term x in the polysemic

In figure 4, displays the term x which is not included in topic 1 or in topic 2. To learn this term included in one topic or both, we use the Wu Palmer algorithm. First, we use some words on the topic to be used as a comparison of the similarity with the term x. Next, we calculate the value of greatest similarity between term x with the term on the topic. If the term x values greater similarity to the term in topic 1. then term x is part of topic 1 and uses vector term on topic 1.

6) Calculate sentence similarity

Finding similarity of words on the topic, we use the cosine similarity algorithm [13] in calculating the value of each word vector as an aspect using the equation:

$$SIM_{(wt,z)} = \frac{\sum_{i=1}^{n} wt_i \times z}{\sqrt{\sum_{i=1}^{n} wt_i^2 \times \sqrt{z^2}}} \quad (6)$$

Where $SIM_{(wt,z)}$ is the value of the similarity between the words w_t to the topic z, z is the vector of the topic as a reference classification document, wt_i is a vector on the word i.

III. RESULT

In this study, we took data testing from Twitter using Twepy crawler in python applications. Data obtained from crawler results on twitter as many as 539 data based on hotel and hospital topics. Each test data will be classified on the topic using cosine equations based on word vectors that have been made with word2vec. In this study to evaluate the proposed method, we use a confusion matrix by looking for precision, recall, and F-measure.

TABEL 3. COMPARISON RESULT OF THE PROPOSED METHOD

Method	Precision	Recall	F-Measure
Word2vec	0.7915	0.9	0.8423
Word2vec + Wu Palmer	0.8232	0.9481	0.8812

The results are shown in table 2 in making the WSD corpus using word2vec have less accuracy. We don't pay attention to new words that are not contained in corpus WSD. The precision value obtained is 0.7915 using word2vec without the addition of a new word. It will be different if the WSD corpus is built with attention to new words that are not in the corpus. To add a new word in the corpus WSD, we use Wu Palmer algorithm to pay attention to meaning between words. Corpus WSD built using word2vec and wu palmer can increase precision values by 0.8232. So that it can improve in classifying the meaning of words on a better topic. When compared without using the addition of a new word on the WSD corpus automatically.

IV. CONCLUSION

From the results of experiments conducted in making corpus for disambiguation by considering the new term that will be added to the corpus. A new term to WSD corpus is added using Word2vec and Wu Palmer. Where word2vec algorithm is used as an initial stage of creating a vector for each term on the topic; then, Wu palmer algorithm is used to add a new term to expand the corpus. The combination of Word2vec and Wu Palmer algorithm can achieve higher precision and recall than those achieved by using only Word2vec to build the corpus WSD.

REFERENCES

[1] S. R. Das and M. Y. Chen, "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web," *Manage. Sci.*, vol. 53, no. 9, pp. 1375–1388, 2007.

[2] K. Dave, K. Dave, S. Lawrence, S. Lawrence, D. M. Pennock, and D. M. Pennock, "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews," *Proc. 12th Int. Conf. World Wide Web*, pp. 519–528, 2003.

[3] E. W. Pamungkas and D. G. P. Putri, "An experimental study of lexicon-based sentiment analysis on Bahasa Indonesia," *Proc. - 2016 6th Int. Annu. Eng. Semin. Ina. 2016*, pp. 28–31, 2017.

[4] B. S. Rintyarna and R. Sarno, "Adapted weighted graph for Word Sense Disambiguation," *2016 4th Int. Conf. Inf. Commun. Technol. ICoICT 2016*, vol. 4, no. c, 2016.

[5] S. Banerjee and T. Pedersen, "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet," *Comput. Linguist. Intell. Text Process.*, vol. 2276, pp. 136–145, 2002.

[6] A. Jain, D. K. Tayal, and S. Vij, "Word sense disambiguation using fuzzy semantic relations," *Proc. 10th INDIACom; 2016 3rd Int. Conf. Comput. Sustain. Glob. Dev. INDIACom 2016*, pp. 2197–2202, 2016.

[7] S. M. Tyar and T. Win, "Jaccard coefficient-based word sense disambiguation using hybrid knowledge resources," *2015 7th Int. Conf. Inf. Technol. Electr. Eng.*, pp. 147–151, 2015.

[8] F. H. Rachman, R. Sarno, and C. Fatichah, "CBE: Corpus-based of emotion for emotion detection in text document," *Proc. - 2016 3rd Int. Conf. Inf. Technol. Comput. Electr. Eng. ICITACEE 2016*, pp. 331–335, 2017.

[9] F. Hastarita Rachman, R. Sarno, and C. Fatichah, "Music Emotion Classification based on Lyrics-Audio using Corpus based Emotion," *Int. J. Electr. Comput. Eng.*, vol. 8, no. 3, pp. 1720–1730, 2018.

[10] E. W. Pamungkas, R. Sarno, and A. Munif, "B-BabelNet: Business-Specific Lexical Database for Improving Semantic Analysis of Business Process Models," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 15, no. 1, p. 407, 2017.

[11] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "502 Distributed Representations of Words and Phrases and Their Compositionality," pp. 1–9.

[12] K. Manjula Shenoy, K. C. Shet, and U. D. Acharya, "A New Similarity Measure for Taxonomy Based on Edge Counting," *Int. J. Web Semant. Technol.*, vol. 3, no. 4, pp. 23–30, 2012.

[13] G. Sidorov, A. Gelbukh, H. Gómez-Adorno, and D. Pinto, "Soft similarity and soft cosine measure: Similarity of features in vector space model," *Comput. y Sist.*, vol. 18, no. 3, pp. 491–504, 2014.

Emotion Label from ANEW dataset for Searching Best Definition from WordNet

Tegar Satrio Utomo
Departement of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
mrsatriotegar@gmail.com

Riyanarto Sarno
Departement of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Suhariyanto
Faculty Of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
Suhariyanto@dsn.dinus.ac.id

***Abstract*—Emotions in a sentence can hold many roles in the field of sentiment analysis, knowing emotions can also be known through facial expressions, speech and text. In this paper we try to find the best definition for WordNet, by utilizing ANEW and ISEAR datasets. Sentences from WordNet that match the parameters of the two datasets will be searched for emotion, labeled according to the class and selected. This can be used in text mining.The result of this research showed that not all definitions in wordnet can accurately match all datasets and contexts**

Keywords—ANEW; Thayer's Model; WordNet; word sense disambiguation; POS tagging

I. INTRODUCTION

Determining emotion in a sentence is one of the studies that many fields apply it, call it text mining, sentiment analysis[14], mood analysis. But if we wrongly define the definition of a word, then the Sentiments and Emotions we seek can be wrong. This time the data we choose is ANEW data and ISEAR data.

Before discussing further, what is ANEW? ANEW [1] is a spatial model of various emotional words, ANEW is a dataset that has many values, one of which is the value of Valence Mean and Arousal Mean. These two values will later be used to find the emotional label of ANEW. Based on the emotional class in the Thayer's model [2] the ANEW word can be searched for the label or emotion class. Then using ISEAR data, will look for examples of sentences to be searched sin for that sentence.

Synset is a collection of synonyms that have the same context, this synset will be related to WordNet, an English lexical[13] database [3]. Sentences in this WordNet which will be seeking the definition of WordNet that matches the emotional label of the word ANEW

II. METHODOLOGY

A. Looking for an emotional label from ANEW data

The first step is to look for the emotional label of each word in the ANEW data. As previously explained that ANEW [1][12] has a Valence Mean value and an Arousal Mean value, these two values contain the datum value of how nice or not a word (valence) and value how calm or happy a word (arousal) [1].

This value will be used to map where the word emotion class in ANEW data, using Thayer's model. [2]

Fig. 1. Thayer's model for emotion classification

Fig. 1 shows what is Thayer's model in general, there are 2 dimensions with 4 classes. Thayer's Model is one of the emotional grouping models commonly used to model emotions in music, Thayer's model [2] has 2 dimensions, in the x-axis and y-axis, for the x-axis the taken quantity is the valence meanwhile for the y-axis taken is the arousal mean.

The value for each axis is a scale of 0 to 9 [4]. ANEW word that will be searched for emotion label then taken valence mean and arousal mean that then mapped into each emotion class in Thayer's Model.

If valence mean < 5 and arousal mean > = 5 then belong to emotion class angry

If valence mean > = 5 and arousal mean > = 5 then belong to emotion class happy

If valence mean < 5 and arousal mean < 5 then belong to emotion class sadness

If valence mean > = 5 and arousal mean < 5 then belong to emotion class calm

B. Matching emotion classes on the Thayer's Model and emotional labels on ISEAR data

The next step is to match the ANEW emotion class and the emotional label from ISEAR. Before discussing further, what exactly ISEAR is? ISEAR [5] is a project run by world psychologists directed by Klaus R. Scherer and Harald Wallbott.

ISEAR collects data from Respondents consisting of psychologists, psychology students, and non-psychologists.

Respondents are asked to give examples of sentences that are real situations in which they have experienced an event that they think has emotion or is an experience consistent with 7 major emotional labels (joy, fear, anger, sadness, disgust, shame, and guilt). [5]

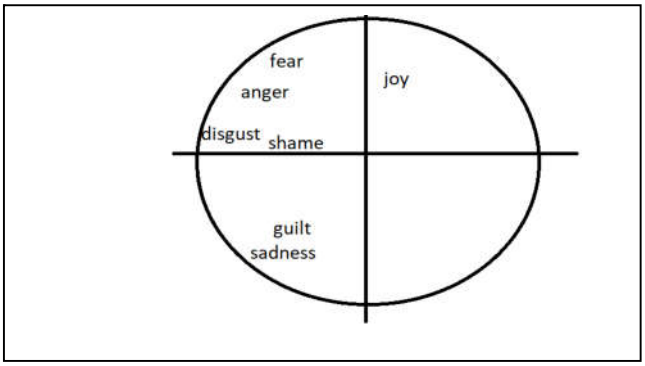


Fig. 2. ISEAR emotion labels

In ISEAR data, emotions have 7 labels, while in Thayer's Model have 4 emotion classes [6], this difference can be uniform by finding a match between the emotion class in the Thayer's Model as well as the emotional label on the ISEAR data. so that each label on ISEAR corresponds to the emotion classes on the Thayer's Model.

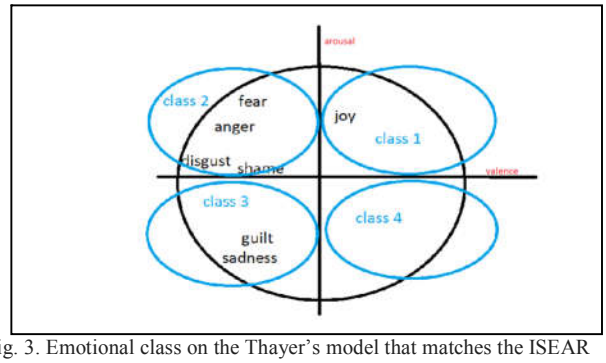


Fig. 3. Emotional class on the Thayer's model that matches the ISEAR emotional label

After Emotional Label on ISEAR is matched on emotion class in Thayer's Model, we get 3 emotion class on Thayer's Model that match with emotional label in ISEAR data, that is class 1 containing joy emotion, class 2 containing anger emotion, fear, disgust, shame class 3 which contains guilt emotion, sadness. Class 4 is not included because there is no match with the sentence's emotion label on the ISEAR data.

Fig. 2 and Fig. 3 shows us about the label and classes that used to process data. Fig. 2 shows the label from ISEAR dataset, there is 7 label in ISEAR that will be grouped to classes that Fig. 1 mention.

C. Looking for sentence examples on ISEAR

As previously known, ANEW data containing words will be sampled sentences in ISEAR Data, the sentence must contain the said ANEW data in question along with an emotion label that is in accordance with the emotion class of the Thayer's Model which is already aligned with the ISEAR data emotion label.

To do string matching, one way that can be used is to use regular expression or regex [7]. By looking for the ANEW word emergence pattern in the sample sentence in the ISEAR data, the example sentence can be retrieved

	Description	Emotion	Sentence
6	guilty	anger	My brother accused me undeservedly before my p...
9	hope	joy	When I was in a situation which seemed hopeles...
17	hospital	joy	I was in hospital, and then was visited by som...
20	hammer	sadness	When I didn't get a certain job at Lillehammer...
27	hug	joy	When I get a hug from someone I love

Fig. 4. Table of ANEW word and its sentences

In Fig.4 the "Description" column contains the word from the ANEW data that has passed the previous step (look for the emotion class). The "Emotion" column contains the ANEW's classified emotion class, and the "Sentence" column contains a sample sentence from the ISEAR data containing the ANEW word and has the appropriate emotion class

C. POS Tagging

After getting a sample sentence from the ISEAR data for every word in ANEW that already has an emotion class, The next thing is to find the word class or POS (Part-of-Speech) tagging of our ANEW word in the sample sentence from the ISEAR data.

Using the nltk library on python. Nltk (Natural Language Toolkit) [8] is a library that provides various features commonly used to process data related to human language, with an interface that is easy to use. some of the features used are WordNet, tokenization, stemming, POS Tagging. And Lesk.

In POS tagging or searching for word classes, some examples of the English word class include:

TABLE I. POS TAGGING IN ENGLISH LANGUAGE

code	class
JJ	Adjective
VB	Verb, base form
NN	Noun
RB	Adverb

a) Noun (N)

If POS tags are starts with N, then it should be changed to 'n'.

b) Verb (V)

If POS tags are starts with V, then it should be changed to 'v'.

c) Adjective (A)

If POS tags are starts with J, then it should be changed to 'a'.

d) Adverb (R)

If POS tags are starts with R, then it should be changed to 'r'.

The goal from this step are to make a structure that called "Synset". So our ANEW structure had a same structure with

Description	Emotion	Sentence	POS
guilty	anger	My brother accused me undeservedly before my p...	JJ
hospital	joy	I was in hospital, and then was visited by som...	NN
hug	joy	When I get a hug from someone I love.	NN
happy	joy	At a friend's birthday party with some of my c...	JJ
hard	joy	At a choral concert I appeared as a solist. Th...	JJ

Wordnet in the form of "Synset" [3].

Fig. 5. POS Tagging

D. Define synset and Define WordNet Definition sentences

Once we get the ANEW emotion labels, get a sample sentence from the corresponding ISEAR data, get the word class from ANEW word in the sample sentence from the ISEAR data, then the next step is to look for synset. As has been explained, synset is a collection of synonyms of words that have the same meaning of the same. When synsets are formed then we can look for examples of sentences that have similar aspects in WordNet.

Synset has the following structure "guilty.a.01 [9]"

The above sentence can be interpreted, The word "guilty" in the sentence has a class adjective and has the closest meaning to the "guilty" word in index 1 in WordNet. To perform a Synset search, one way is to use a latch algorithm [10]. how does it work lesk? The way the lesk works is to think that the adjacent word is a similar word.

Lesk returns a synset value of the ambiguous word and the context in which it appears.

```
>>> from nltk.wsd import lesk
>>> sent = ['I', 'went', 'to', 'the', 'bank', 'to', 'deposit', 'money', '.']

>>> print(lesk(sent, 'bank'))
Synset('savings_bank.n.02')
```

Fig. 6. The Lesk Algorithm for searching Synset

In the words ANEW as ambiguous word and sentence from ISEAR as word exit context, as well as POS tagging from ANEW word, Lesk returns Synset with the highest number of word overlapping between different benchmarks and definitions of each Synset.

So one word from WordNet will be chosen which has aspects and uses that are similar to the words from ANEW.

III. EXPERIMENT

To prove the correctness of the chosen definition, the ISEAR sentence example is searched for by measuring how much the number of words ANEW in the sentence [11]

$$v_{text} = \frac{\sum_{i=1}^{n} v_i f_i}{\sum_{i=1}^{n} f_i} \quad (1)$$

where :

v_i = Average valence for word *i*

f_i = frequency for ANEW word

To estimate the overall valence score for a text, which we denote by $\mathbf{V}_{\text{text}}$, we

(a) determine the frequency f_I that the ith word from the ANEW study word list appears in the text; and

(b) compute a weighted average of the valence of the ANEW study words as shown in equation (1).

After emerging Emotion results from the ISEAR sentence, then we calculate how accurate we calculated.

To calculate the accuracy of the confusion matrix, The accuracy is simply the sum of the diagonal elements divided by the sum of all elements in the matrix.

TABLE II. THE EXPERIMENT RESULTS

Predicted / Actual	anger	joy	Sadness	other	NaN
Anger	25	8	6	0	8
Joy	1	50	1	2	24
Sadness	4	4	9	1	5

After the calculation process from 148 data found, there are 84 data that match the label,

then we use the following formula to find accuracy

$$\text{Accuracy} = \frac{tp + tn}{tp + fn + fp + tn} \quad (2)$$

where :
tp = true positive
tn = true negative
fp = false positive
fn = false negative

So the accuracy value is 0.56, with an accuracy value of 0.56 there are still many developments that can be done to increase the value of accuracy. However, the definition of WordNet that we take is right for every word ANEW.

IV. CONCLUSION

By utilizing the Synset structure formed by the two datasets that we have several method that we used, we can get WordNet definition that fits the context we want and measure how accurately we determine it.

ACKNOWLEDGEMENT

Authors would like to thank Institut Teknologi Sepuluh Nopember and the Ministry of Research, Technology, and Higher Education of Indonesia for supporting the research.

V. REFERENCES

[1] M. M. Bradley and P. J. Lang, "Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings NIMH Center for the Study of Emotion and Attention Affective Norms for English Words (ANEW) 1999."

[2] J. Grekow and Z. W. Raś, "Detecting Emotions in Classical Music from MIDI Files," Springer, Berlin, Heidelberg, 2009, pp. 261–270.

[3] A. Javed and B. S. Lee, "Hybrid semantic clustering of hashtags," *Online Soc. Networks Media*, vol. 5, pp. 23–36, Mar. 2018.

[4] R. E. Thayer, *The biopsychology of mood and arousal*. Oxford University Press, 1989.

[5] H. G. Wallbott and K. R. Scherer, "How universal and specific is emotional experience? Evidence from 27 countries on five continents," *Soc. Sci. Inf.*, vol. 25, no. 4, pp. 763–795, Dec. 1986.

[6] F. H. Rachman, R. Sarno, and C. Fatichah, "Music Emotion Classification based on Lyrics-Audio using Corpus based Emotion," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 3, pp. 1720-1730., Jun. 2018.

[7] C. Chapman, P. Wang, and K. T. Stolee, "Exploring regular expression comprehension," in *2017 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE)*, 2017, pp. 405–416.

[8] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, (2009)

[9] D. Jurafsky and J. H. Martin, *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech*

[10] B. S. Rintyarna and R. Sarno, "Adapted weighted graph for Word Sense Disambiguation," in *2016 4th International Conference on Information and Communication Technology (ICoICT)*, 2016, pp. 1–5.

[11] P. S. Dodds and C. M. Danforth, "Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents," *J. Happiness Stud.*, vol. 11, no. 4, pp. 441–456, Aug. 2010.

[12] F. H. Rachman, R. Sarno, and C. Fatichah, "CBE: Corpus-based of emotion for emotion detection in text document," in *2016 3rd International Conference on Information Technology, Computer, and Electrical Engineering (ICITACEE)*, 2016, pp. 331–335.

[13] E. W. Pamungkas, R. Sarno, and A. Munif, "B-BabelNet: Business-Specific Lexical Database for Improving Semantic Analysis of Business Process Models," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 15, no. 1, p. 407, Mar. 2017.

[14] B. S. Rintyarna, R. Sarno, and C. Fatichah, "Enhancing the performance of sentiment analysis task on product reviews by handling both local and global context," *International Journal of Information and Decision Sciences*. 2018 Vol. 11.

Searching Word Definitions in WordNet Based on ANEW Emotion Labels

Pius Pambudi
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
piuspbd@gmail.com

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Edi Faisal
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
faisal@dsn.dinus.ac.id

***Abstract*—Emotion detection is the process of extracting emotion from human expression. Emotions may be expressed by facial expression, speech and even written text. The Affective Norms for English Words (ANEW) is a corpus that has been used in researches involving emotion detection. ANEW provides the score of valence, arousal, and dominance for each word. However, no information about the word definition is provided in ANEW. The goal of this research is to identify the correct definition of every word in ANEW as additional informations to improve its accuracy.**

Keywords—emotion; affective norms; word sense disambiguation;

I. INTRODUCTION

The Affective Norms for English Words (ANEW) is a collection of words that developed by Margaret M. Bradley & Peter J. Lang in 1999 [1]. ANEW has 1030 words and their emotion components, such as valence, arousal, and dominance value. ANEW has been used to calculate emotion in song titles, lyrics and even weblogs [2].

Affective Norms for English Words (ANEW) is the dataset that is often used for research on the dimensional model [3]. The dimensional model describes emotions in a dimensional vector space [4]. In ANEW itself, there are three-dimensional values: valance, arousal and dominance. Valence and arousal represent personal emotions, while dominance shows the relationship between people and their environment. In vector space, the range of valence value from negative into positive and arousal has a value from low until high [4]. Although most research focuses only on one-dimensional model [5], here this research will use two-dimensional model of emotions: valence and arousal.

Emotion detection of text has been considered important in analyzing personal emotion [6]. While ANEW has been used in a lot of these research, an improvement still can be made to improve its accuracy. For example, at the moment, ANEW only contains words without information about their senses and definitions. This can be a problem because some words actually have multiple senses and definitions. When using ANEW, we can potentially use the wrong emotion value of a word in ANEW that has completely different meaning in our context even though they appear as the same word.

The idea of this paper is to narrow down the definition of each ANEW word so that such problem cannot happen.

II. METHODOLOGY

A. Labelling

Firstly, we will need to add an emotion label to each word in ANEW. This can be achieved by calculating their valence and arousal means. There are a lot of kinds of emotion classification [7]. One of them is known as Thayer's model [8].

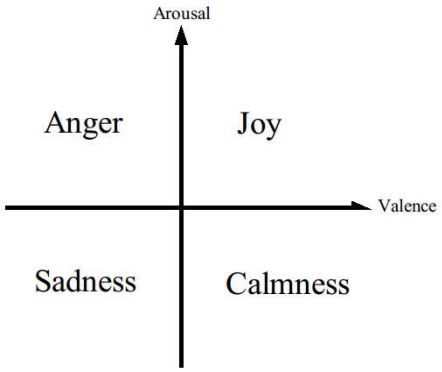
This experiment uses Thayer's model for the emotion labels because it has emotion classes with clear limits on dimension spaces [8]. Thayer's model supports four classes of emotion (joy, anger, sadness, calmness), but this experiment will only use three of them (joy, anger, sadness) due to dataset availability.

The table below shows the examples of word in ANEW. While, the figure below shows how Thayer's model in a diagram.

TABLE 1. Affective Norms for English Words (ANEW)

	Description	Valence	Arousal
	grin	7.40	5.27
	honest	7.70	5.32
	gripe	3.14	5.00
	honey	6.73	4.51
	guillotine	2.48	6.56

FIGURE 1. Thayer's model



B. Dataset

Dataset is gathered from ISEAR Databank (International Survey on Emotion Antecedents and Reactions), a project directed by Klaus R. Scherer and Harald Wallbott in 1997. ISEAR is a dataset that contains sentences and their respective emotion labels. Out of seven major emotions (joy, fear, anger, sadness, disgust, shame, guilt) in ISEAR, this experiment will use only three of them (joy, anger, sadness). This dataset will later be used as sample sentences to add contexts to our ANEW words before we can find the definitions.

After giving each ANEW word an emotion label, we can add sentences from this dataset to our ANEW words. For each word in ANEW, we will attach a sentence from ISEAR that contains the word and has the same emotion label.

It is important to note that not all words in ANEW will have sample sentences from ISEAR. In later steps, we will only use words that have sample sentences.

Here is some examples of ISEAR sentences.

TABLE 2. ISEAR databank

	Sentence	Label
	During the period of falling in love, each time that we met and especially when we had not met for a long time.	joy
	When I was involved in a traffic accident.	fear
	When I was driving home after several days of hard work, there was a motorist ahead of me who was driving at 50 km/hour and refused, despite his low speed to let me overtake.	anger
	When I lost the person who meant the most to me.	sadness
	The time I knocked a deer down - the sight of the animal's injuries and helplessness. The realization that the animal was so badly hurt that it had to be put down, and when the animal screamed at the moment of death.	disgust

C. POS Tagging

For each sample sentence, we will need to find POS of each word in the sentence. By doing this, we can get the POS of all ANEW words in each of their sample sentences.

POS is important so that we can narrow the word definition by only searching definitions that match the POS and ignoring the rest.

D. Find Word Definition with WordNet

WordNet is a lexical semantic database for the English language [9]. WordNet consists of the definitions for all English words and their relation called *synset*. With WordNet, we can search for all definitions for each word in ANEW with the POS tags that we have calculated earlier. However, WordNet usually will match multiple definitions for each word.

At this point, we must choose one from these definitions which match our sample sentence the best. Such problem is called Word Sense Disambiguation (WSD). Word Sense Disambiguation is a problem to remove ambiguity and to find the meaning of words in a particular context [10]. One example of a WSD algorithm is called Lesk algorithm.

Lesk algorithm is a classical WSD algorithm introduced by Michael E. Lesk in 1986 [11]. A simplified version of Lesk algorithm has been adapted in WordNet and can be used to find the definition of an ambiguous word [12]. Lesk algorithm will calculate the similarity between our sentences and WordNet definitions and will result in one definition that has the best similarity to our sentence.

In the end, we will have one definition for each word in ANEW.

E. Testing

We need to test our methods and calculate the accuracy. A simple testing method would be to gather the definition and all sentence examples (from WordNet) that we have obtained earlier and then for each word in that sentence, if the word exists in ANEW, we calculate its valence and arousal value and then calculate the average.

Here is the formula to calculate the average valence and arousal of a text. Here, v_i and a_i denote respectively the valence and arousal value for each word in ANEW. f_i denotes the frequency of occurrence of those words in the text.

$$v_{text} = \frac{\sum_{i=1}^{n} v_i f_i}{\sum_{i=1}^{n} f_i} \quad v_{text} = \frac{\sum_{i=1}^{n} v_i f_i}{\sum_{i=1}^{n} f_i} \tag{1}$$

$$a_{text} = \frac{\sum_{i=1}^{n} a_i f_i}{\sum_{i=1}^{n} f_i} \quad a_{text} = \frac{\sum_{i=1}^{n} a_i f_i}{\sum_{i=1}^{n} f_i} \tag{2}$$

The resulting valence and arousal average then can be used to find the emotion label based on Thayer's model.

Lastly, we need to compare these emotion labels with the emotion labels that we have calculated in the first method earlier.

III. EXPERIMENT

Initially, we used all 1030 words in ANEW and labelled them with emotion labels (joy, anger, sadness). Since we only used three emotion labels from supposedly four, we had a few words that did not have labels.

TABLE 3. Emotion labelling

Valence	Arousal	Emotion
>= 5	>= 5	joy
< 5	>= 5	anger
< 5	< 5	sadness

TABLE 4. Emotion labelling on ANEW

	Description	Valence	Arousal	Emotion
	grin	7.40	5.27	joy
	honest	7.70	5.32	joy
	gripe	3.14	5.00	anger
	honey	6.73	4.51	-
	guillotine	2.48	6.56	anger

After finding sample sentences for each ANEW word. We ended up with 189 words that had sample sentences and were ready for the next step.

TABLE 5. Adding sample sentences

	Description	Emotion	Sentence
	guilty	anger	A chief was unjustly punished on an autumn bri...
	hope	joy	My friend took me to Civo Stadium to watch and...
	hospital	joy	I felt it when on duty in the hospital I under...
	hug	joy	When I get a hug from someone I love.
	happy	joy	At a friend's birthday party with some of my c...

Using sample sentences we gathered, we could find the POS tag of each ANEW word in its respective sentence.

TABLE 6. POS tagging

	Description	Emotion	Sentence	POS
	guilty	anger	A chief was unjustly punished on an autumn bri...	JJ
	hope	joy	My friend took me to Civo Stadium to watch and...	NN
	hospital	joy	I felt it when on duty in the hospital I under...	NN
	hug	joy	When I get a hug from someone I love.	NN
	happy	joy	At a friend's birthday party with some of my c...	JJ

Afterwards, we attempted to get WordNet *synset* for every word. However, when finding POS tags earlier, there was a case when WordNet failed to find a word definition that matched the POS tag. This could happen because of the incorrect POS tag. The total number of dataset decreased to 171 words.

TABLE 7. Obtaining definitions

	Description	Emotion	POS	Synset
	guilty	anger	JJ	guilty.a.01
	hope	joy	NN	promise.n.02
	hospital	joy	NN	hospital.n.02
	hug	joy	NN	hug.n.01
	happy	joy	JJ	happy.a.01

Finally, we did the testing process using Formula (1) and (2) we defined earlier. The test resulted in a new emotion label for every word. By comparing these results with the initial labels, we could conclude the correctness of this method.

However, in the testing process, there was also a case when the word definition and its WordNet examples were too brief to test against the words in ANEW. We excluded these instances in further steps.

TABLE 8. Emotion testing

	Description	Emotion label	Test result	Verdict
	guilty	anger	anger	correct
	hope	joy	joy	correct
	hospital	joy	anger	incorrect
	hug	joy	joy	correct
	happy	joy	joy	correct

In the end, we counted how many instances were predicted correctly. Out of 149 words that have been successfully tested, 99 of them resulted in a correct emotion label. We concluded that the final accuracy was 99/149 = 66.44%.

IV. CONCLUSION

Our experiment has successfully tested 149 words and predicted the correct emotion labels for 66.44% of our dataset.

The word definitions could be used as additional informations when using ANEW in a research with the hope of better accuracy.

However, there are still rooms for improvement for this research to achieve the more accurate definition of every word in ANEW.

V. REFERENCES

[1] M.M. Bradley and P.J. Lang, "Affective Norms for English Words (ANEW): Instruction Manual and Affective Ratings". *Technical Report C-1, The Center for Research in Psychophysiology, University of Florida*, 2008.

[2] P.S. Dodds and C.M. Danforth, *"*Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents". *Journal of Happiness Studies,* pp 441–456, 2010.

[3] H. Corona and M.P.O. Mahony, "An Exploration of Mood Classification in the Million Songs Dataset", *12th Sound and Music Computing Conference*, 2015.

[4] F.H. Rachman, R, Sarno and C. Fatichah, *"*Music Emotion Classification based on Lyrics-Audio using Corpus based Emotion". *International Journal of Electrical and Computer Engineering (IJECE),* Vol 8, No 3, pp 1720-1730, 2018. DOI: http://doi.org/10.11591/ijece.v8i3.pp%25p.

[5] Y. Hu, X. Chen, and D. Yang, "Lyric-Based Song Emotion Detection with Affective Lexicon and Fuzzy Clustering Method", *ISMIR 2009*, pp. 123-128, 2009.

[6] F.H. Rachman, R. Sarno and C. Fatichah, *"*CBE: Corpus-based of emotion for emotion detection in text document". *Information Technology, Computer, and Electrical* Engineering (ICITACEE), 2016. DOI: 10.1109/ICITACEE.2016.7892466.

[7] B. Thomas, P. Vinod, and K.A. Dhanya. "Multiclass Emotion Extraction from Sentences". *International Journal of Scientific & Engineering Research*, Vol 5, Issue 2, 2014.

[8] R.E. Thayer, "The Biopsychology of Mood and Arousal", *New York: Oxford University Press*, 1989.

[9] C. Fellbaum, "WordNet: An Electronic Lexical Database", *MIT Press*. 1998.

[10] B.S. Rintyarna and R. Sarno, *"*Adapted weighted graph for Word Sense Disambiguation". *Information and Communication Technology (ICoICT),* 2016. DOI: 10.1109/ICoICT.2016.7571884.

[11] M. Lesk. "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone." *SIGDOC '86: Proceedings of the 5th annual international conference on Systems documentation*, pp 24-26. 1986.

[12] S. Banerjee and T. Pedersen. "An Adapted Lesk Algorithm for Word Sense Disambiguation Using WordNet", *Lecture Notes in Computer Science*, Vol. 2276, Pages: 136 - 145, 2002.

Digital Literacy Skills of Informatics Engineering Education Students as The Basis for Online Learning Implementation

Hakkun Elmunsyah
Electrical Engineering Department
Universitas Negeri Malang
Malang, Indonesia
hakkun@um.ac.id

Wahyu Nur Hidayat
Electrical Engineering Department
Universitas Negeri Malang
Malang, Indonesia
wahyu.nur.ft@um.ac.id

Syaad Patmantara
Electrical Engineering Department
Universitas Negeri Malang
Malang, Indonesia
syaad.ft@um.ac.id

***Abstract*— Learning in the century of knowledge means that each individual can learn anywhere, anytime, with anyone, and through any media. The existing of technological developments are also ready to support the implementation of these online learning in desktop, web and mobile. With optimal utilization of ICT, it is hoped that learning can take place more attractive, real, and fun. The next problem that needs to be prepared is the assurance of digital literacy skills learners as the basis of the online learning application. Therefore, this research will be studied to the digital literacy skills of informatics engineering education students at Universitas Negeri Malang which are focused on the ability of information literacy, media literacy, and computer literacy. The research method is descriptive analysis with 99 respondents. The results of this study shows that the level of information literacy and media literacy students are in good category with the average value of 74,25% and 72,75%, while tech literacy ability is in very good category with the average value 83,75%. Based on these results it can be concluded that IT students have good literacy skills and can be improved through online and web-based learning.**

Keywords— digital literacy skills, online learning, ICT

I. INTRODUCTION

The pattern of human life is always changing due to the development of science and technology. From time to time, more and more technologies are being created to add to the comfort of work and human life. Technological advances also penetrated into the world of education that causes various behavioral changes in learning and teaching activities, learning resources, and learning technology. Some changes in learning behavior, among others, caused by technological developments in the era of industry 4.0 which causes connectedness with the global world more easily and affordably.

Industrial world 4.0 based on the concepts and technologies that include cyber-physical systems, the Internet of things (IoT), and the Internet of services (IoS) [1][2][3], based on perpetual communication via Internet that allows a continuous interaction and exchange of information not only between humans (C2C) and human and machine (C2M) but also between the machines themselves (M2M)[4]. This communicational interaction influences the establishment of knowledge management 4.0 (KM 4.0) [5]. It is closely related with the Internet of Things (IoT), Cyber Physical System (CPS), information and communications technology (ICT), Enterprise Architecture (EA), and Enterprise Integration (EI) [6]. The industrial revolution 4.0 had an impact on the reduction of the role of human labor that was replaced by machines. It is a challenge in the global information age that demands the world of higher education to prepare competent and competent human resources to meet the demanding needs of society and industry along with the rapid development of science and information technology [7].

Characteristics of the world of work and the qualifications of labor that the industry needs in the industry era 4.0 is changing rapidly [8]. One of the industry's traits of the century is the increasing need for the generic skills attributes that workers must possess [9]. Other survey results show that employment recruitment by firms prioritizes employability skills / soft skills rather than hard skills [10]. Therefore, work-oriented education through the mastery of technical skills and employability skills is needed to sustain economic development in the XII century.

Changes in the paradigm that occurs today requires every individual to be able to adapt to the times. Especially in the days that have been digitized as it is today. Various information, data, and knowledge of both fact and opinion, can be easily obtained. Therefore, the ability to select, process, and interpret the acquired digital information is essential to the mastery of the students. These skills are called digital literacy skills [11], [12], [13]. Digital literacy skills are more than just the ability to search for information by using digital devices properly; it consists of the skills to build a knowledge in a complex and systematic way such as browsing, then interpreting data and information, creating and sharing content on the web, chatting via chat rooms and communicating in social networks [14].

On the other hand, ICT advances affect the use of teaching aids in schools and other educational institutions. The impact of ICT advancement on learning in schools is by changing and shifting educational paradigms. In order to improve the effectiveness and efficiency of learning, it is necessary to develop various creative and innovative learning models. This needs to be done so that the learning process does not seem less interesting, monotonous and boring so that will inhibit the transfer of knowledge. Learning media is a very important thing in the world of education used as a learning facility for students. Learning media can help students to understand and master the material learned. Learning media that apply the concept of interactive multimedia aimed at improving students' motivation in learning. With the learning media is expected to improve the effectiveness of learning and the desire of students to learn more independently.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

With the implementation of ICT learning in the classroom, whether in the form of full online learning or blended learning digital literacy capability is required good [15]. The ability of digital literacy is the basic ability students have to have when learning is done online. With this ability, students can sort and choose well the material that suits their needs and quality. This means that with the ability of digital literacy is good then the learning process will be more optimal. Especially in the era of generation Z (digital native), students face a world that is completely open and connected in all lines. What new things happen in other parts of the world, can be quickly obtained information. Therefore, as an important basic capability in the application of online learning, in this paper will be discussed about the readiness of the implementation of online learning seen from the level of digital literacy ability of students.

II. METHOD

A. Data Collection

This research is a quantitative descriptive research conducted by distributing questionnaires to respondents. The population of the research is the student of informatics engineering education of Universitas Negeri Malang with the number of research sample is 99 students. Students of informatics techniques are identical with their habits in utilizing ICT, therefore in this study used as respondents to know the level of digital literacy.

B. Digital Literacy Skills Measurement Tools

For digital literacy skills instrument in this research, used questionnaire with likert scale. Likert scale is chosen because it can be used to measure attitudes, opinions, and perceptions of a person or group about social events or symptoms. The measurement model used is Likert Scale with modification of four answer options. Alternative answers used are Strongly Agree, Agree, Disagree, and Strongly Disagree. The statement items are presented in two forms: positive revelation and negative statements. A positive statement is a statement that supports negative ideas and statements is a statement that does not support the idea.

The instrument of digital literacy capability is developed from 3 aspects [16], namely (a) literacy or information literacy, which among other things contain reading ability in a broad sense (not only reading text but also "reading pictures") and understanding what is read adequately. Operational students are able to access information effectively (source nformasi) and efficient (time), evaluate information to be used critically and competently, and use and manage information accurately and efektf to solve problems; (b) media literacy, which means the skills of utilizing various media, both electronic and non-electronic in order to obtain extensive information and pemahan. students are able to select and develop the medium used to communicate; and (c) technology literacy where students are able to use and create appropriate media for communication.

The instrument of digital literacy capability developed has been through content validation and construct validation process. Content validation is done by consultation of grid with instrument which has been developed to 21st century learning expert. Based on validation result of content which has been done by expert judgment then continued revision and tested to student outside research sample. This validation is called construct validation with the number of respondents as many as 100 students. The validation result states that the overall instrument of information literacy (30 items), media literacy (20 items), and technological literacy (37) are entirely valid. For all instrument, its purpose, and value of reliability can be seen in Table 1.

Table 1. Digital Literacy Skills Measurement Tools

Aspect	Instruments Purposes	Items (n)	Reliability Score
Information literacy	Measure perception of ability to select and understand information	30	0,934
Media literacy	Measure perception of skills utilizing various media	20	0,895
Technology literacy	Measure perception of skills using technology	37	0,956

C. Analysis Method

The analytical method used is descriptive analysis, where the score item of each indicator will be averaged. The result of the average score is then interpreted in accordance with the data interpretation table. Interpretation of meaning in the study will use four categories / levels, namely: (1) very good, (2) good, (3) less good, and (4) is not good. Based on the calculation obtained the result of the ideal score range per item / item of questionnaire in the category of very good, good, bad, and very bad as written in Table 2.

Table 2. Data Interpretation Category

No	Interval	Percentage	Category
1	3,25 – 4,00	81,25 - 100	Very good
2	2,50 – <3,25	62,50 - <81,25	Good
3	1,75 – <2,50	43,75 - <62,50	Bad
4	1,00 – <1,75	25,00 - <43,75	Very bad

III. RESULT AND DISCUSSION

Literacy can be interpreted as a reading and writing ability. Literacy is also known as literacy or literacy, but nowadays literation has a broad meaning, so literacy is no longer meaningless but meaningful. Variety of literacy such as computer literacy, media literacy, technological literacy, economic literacy, information literacy, and even moral literacy. Literacy or literacy can be understood literate technology, information literacy, critical thinking, sensitive to the environment, even politically sensitive. A someone is said to be literate if it is able to understand something because it reads the right information and does something based on its understanding of the content of the page. In the study of digital literacy capabilities translated into three aspects namely the ability of information literacy, media literacy, and technology literacy (computer).

A. Information Literacy

Literacy or information literacy means the ability to read in a broad sense (not just reading the text but also "reading the picture) and understanding what is read adequately. Operational students are able to access information effectively (source of information) and efficient (time), evaluate the information to be used critically and competently, and use and manage information accurately and effective to overcome the problem.

Table 3 Results of Information Literacy Analysis

No	Aspect	Mean	%	Category
1	Identification	2,82	70,50	Good
2	Eksploration	3,04	76,00	Good
3	Selection	3,00	75,00	Good
4	Organization	3,10	77,50	Good
5	Creating	2,95	73,75	Good
6	Presentation	2,80	70,00	Good
7	Evaluate	2,93	73,25	Good
8	Application	3,08	77,00	Good
	Mean Score	2,97	74,25	Good

Based on Table 3 it can be seen that the average score literasi information entered in either category. This means that the ability of students in managing and understanding the information is good. With this ability, students are expected to sort and select information that suits their needs, including distinguishing the hoax information or not.

The ability of information literacy is critical to the success of lifelong learning, information literacy is a major competence in the information age and information literacy contributes to the development of teaching and learning. The term information literacy is always associated with computer literacy, library skills and critical thinking which is a support for the development of information literacy [17].

B. Media Literacy

Media literacy which means skills utilize various media, both electronic and non-electronic in order to obtain information and broad pemahan. Students are able to select and develop the medium used to communicate and expand knowledge. The student's digital literacy capability is shown in Table 4.

Table 4 Results of Media Literacy Analysis

No	Aspect	Mean	%	Category
1	Media Type	2,71	67,75	Good
2	Media Function	3,31	82,75	Very Good
3	Content analysis of media messages	2,64	66,00	Good
4	Content evaluation of media messages	3,04	76,00	Good
5	Create Media message contens	2,85	71,25	Good
	Mean Score	2,91	72,75	Good

Table 4 shows that students' media literacy skills fall into either category. It has the meaning that students have the ability to understand, analyze, access, and produce good mass communication messages. Media literacy is understood as a process of reading media content with the ability to analyze, evaluate, and construct messages in a wide and varied form.

C. Technology Literacy

Literacy technology (computer) is the ability to use computers to satisfy personal needs. Another meaning of computer literacy is "A collection of abilities, knowledge, understanding, values, and relationships allowed by a person to provide a comfortable function as a productive society using a community-oriented computer.

Table 5 Results of Technology Literacy Analysis

No	Aspect	Mean	%	Category
1	Word processing application	3,44	86,00	Very Good
2	Number Processor	3,02	75,50	Good
3	Media presentation	3,38	84,50	Very Good
4	Use basic computer	3,42	85,50	Very Good
5	Use of network and internet	3,19	79,75	Good
	Mean Score	3,29	83,75	Very Good

Table 5 shows that the technological literacy ability of the students of informatics engineering education is very good. This is certainly influenced by the habit of using the computer, such us when they use blended learning in study [18]. Technology literacy is limited to computer literacy that is a collection of skills, knowledge of understanding, values in communication by using computers to meet personal needs.

In general, digital literacy capabilities are influenced by previously embedded knowledge. Such knowledge can guide users in understanding digital navigation patterns, such as home, site, or travel [19][20][21]. In this study, the metaphorical model uses the English context only. Idealized Cognitive Models (ICM) form the coherence between knowledge and social experience in the cyber genre environment [22].

Cyber genre is understood by the provision of digital literacy capabilities. This understanding is a cognitive process in which users interpret the digital environment with their embedded knowledge and its application in life. For example, the expression "home" or "visit" is used in a new digital context based on the projection or mapping from the social domain to the digital domain.

In the century of knowledge, Z-generation students provided better multitasking skills and were more productive than previous generations [23][24]. This is because, nowadays students are very easy to access information and social media from an early age, so that the Z generation has a good ability to process much of the information they collect [25]. In addition, Z genaration also has characteristics with prominent self-confidence, high initiative, and more creative and innovative to create new opportunities [26].

Z-generation students are very capable in utilizing digital media to support the construction of knowledge [27]. With students' digital literacy skills, students will have creative, critical thinking and views, so they will not be easily ingested by provocative issues, so victims of hoax or digital-based fraud. Digital literacy skills are not just the ability to locate and manage information, but also research and integrate that information [28]. According to Nelson, Courier, & Joseph [29] digital literacy skills can be seen from the skills and knowledge to use the application of digital media software and hardware; the ability to comprehensively understand digital media content; and the peak is the ability to use digital technology creatively to produce work. Therefore, in today's digital age, the improvement of digital literacy skills of students must be continuously pursued by diversifying various models and instructional media, such us implementation of blended learning or e-learning.

IV. CONCLUSION

The discussion of digital literacy capabilities as a basis for the ability that must be possessed for the application of online learning has been discussed in this paper. Digital literacy capability is focused on 3 capabilities, namely information literacy, media literacy, and technology literacy.

The result of the research shows that the level of information literacy and media literacy of UM technique of Informatics students are in good category with the average value of 74,25% and 72,75%, while tech literacy ability is in very good category with the average value 83,75%. This good result is influenced by the characteristics of Z-generation students who have high confidence, multitasking, innovative, creative, and familiar with the digital world, so that when online learning, students easy to process, analyze, and interpret the information provided. With good digital literacy skills, students are expected to have a critical-creative mindset and outlook, so it is not easily affected by negative information by validating the truth of the information.

Referring to the findings of this study it is suggested that: (1) lecturers should apply ICT-based learning to optimize learning; (2) lecturers can diversify media and methods to make learning implementation more attractive, innovative, interesting, and fun especially by utilizing the development of ICT; and (3) lecturers should improve their ability, especially in the field of ICT that can be implemented in learning.

REFERENCES

[1] V. Roblek, M. Meško, and A. Krapež, "A complex view of industry 4.0," SAGE Open, April-June: 1-11, 2016. doi: 10.1177/2158244016653987

[2] Ning, H., & Liu, H "Cyber-physical-social-thinking space based science and technology framework for the Internet of things". Science China Information Sciences, 58, 1-19, 2015. doi:10.1007/s11432-014-5209-2

[3] Lasi, H., Fettke, P., Kemper, H. G., Feld, T., and Hoffmann, M.. Industry 4.0. Business & Information Systems Engineering, 6,

239-242, 2014. doi:10.1007/s12599-014-0334-4

[4] Cooper, J., & James, A. "Challenges for database management in the Internet of things". IETE Technical Review, 26, 320-329, 2009. doi:10.4103/0256-4602.55275

[5] Dominici, G., Roblek, V., Abbate, T., & Tani, M, "Click and drive": Consumer attitude to product development. Towards future transformations of driving experience. Business Process Management Journal, 22, 420-434, 2016. doi:10.1108/BPMJ-05-2015-0076

[6] Y. Lu, "Industry 4.0: A survey on technologies, applications and open research issues", Journal of Industrial Information Integration, Vol. 6, pp:1-10, Juni 2017. doi: 10.1016/j.jii.2017.04.005

[7] Lattuca, Terenzini and Volkwein. "Engineering Change: a study of the impact of EC2000," Baltimore: ABET Inc, 2006.

[8] Tome, E., "Employability, skills, and training in portugal (1988-2000): evidance from official data.," Journal of European Industrial Training, vol.31:5, pp:336-357, 2007.

[9] Gibb, J "Generic skills in vocational education and training," Adelaide SA: National Cetre for Vocational Education Research Ltd, 2004.

[10] Heimler, R. "Attitudes of college graduates, faculty, and employers regarding the importance of skills acquired in college and needed for job performance and career advancement potential in the retail sector (Ed.D. dissertation). Dowling College, United States - New York. Available from ProQuest Dissertations & Theses: Full Text. 2010 (UMI No. 3406253).

[11] Hargittai, E. and Hinnart, A, "Digital inequality: differences in young adults use of the internet," Communication Research, vol. 35-5, pp:602-621, 2008.

[12] Lankshear, C., and Knobel, M. "Digital literacies: concepts, policies and practices," New York: Peer Lang Publishing, 2008.

[13] Eshet, Y., "Thinking in the digital era: a revised model for digital literacy," Issues in Informing Science and Information Technology, vol:9, pp:267-276, 2012

[14] Kavalier J., and B., Flannigan, S, "Connecting the digital dots: literacy of the 21st century," Educause Quarterly, vol: 29-2, pp:1-3, 2006.

[15] S. Patmanthara and W. N. Hidayat, "The Effectiveness of Learning Management System (LMS) on Computer Assisted Learning Course for Informatics Engineering Education Students," Advanced Science Letters, vol. 24, no. 4, 2018, pp. 2642-2645.

[16] Trilling, B. and Fadel, C, "21st century skills: learning for life in our times," San Francisco: John Wiley & Sons, 2009.

[17] Chan S.N. and Z. Ramly, "Information literacy course(GXEX 1401); a survey of it's effectiveness as perceived by students". Kekal Abadi , vol. 20-1, pp:1–7, 2001.

[18] S. Patmanthara and W. N. Hidayat, "Improving Vocational High School Students Digital Literacy Skill through Blended Learning Model,"Journal of Physics: Conference Series, vol. 1028, no. 1, 2018, pp. 1-7 (doi :10.1088/1742-6596/1028/1/012076)

[19] Navarro, I. "Metaphorical ICMs in cybergenre representation," vol 7, Studia Universitatis Petru Maior. Series Philologia, Targu Mures. pp:167-177, 2008.

[20] Navarro, I. and A. J. Silvestre. "The role of icms in cybergenre representation and reading modes," In M. Navarro Coy (ed.) Practical Approaches to Foreign Language Teaching and Learning. Linguistic Insights 78. Berlin/NY: Peter Lang. pp:269-292, 2009.

[21] Navarro, I., M. L. Villanueva, C. Girón-García dan A. J. Silvestre. "Cybergenres and autonomous language learning: pragmatic strategies and cognitive models in the production and processing of digital genres," INTED2008 Proceedings, International Association for Technology, Education and Development, IATED Publications. Pp:777-786, 2008.

[22] Lakoff, G. "Women, fire and dangerous things. what categories reveal about the mind," The University of Chicago Press. 1987. DOI: 10.7208/chicago/9780226471013.001.0001

[23] Ozkan, M., & Solmaz, B. "Mobile addiction of generation z and its effects on their social lifes," Procedia-Social and Behavioral Sciences, vol.205, pp:92-98, 2015

[24] Iorgulescu, Maria C "Generation Z and its perception of work". Cross-Cultural Management Journal, Vol.XVIII-9, pp:47-54, 2016

[25] Bolser, K. & Gosciej, R. "Milennials: multi-generation leaders staying connected". Journal of Practical Consulting, Vol.5-2, pp:1-9, 2015.

[26] Singh, Anjali. "Challenges and issues of generation Z" IOSR Journal of Business and Management, vol. 16(7), pp:59-63, 2014.

[27] Saliyeva, et al. "On the Students' Ability to Use Digital Educational Resources". International Journal Of Environmental & Science Education, vol. 11(11), pp:4669-4679, 2016.

[28] Greene, J.A., Yu, S.B. and Copeland, D.Z. "Measuring Critical Components of Digital Literacy and their Relationships with Learning". Computers & Education, Vol.76, pp:55-69, 2014.

[29] Nelson, KM, Courier, M and Joseph, GW. "Teaching tip: an investigation of digital literacy needs of students". Journal of Computer Assisted Learning, Vol. 28(6), pp:547-556, 2011.

Sentiment Analysis of Indonesian News Using Deep Learning
(Case Study : TVKU Broadcast)

[1]Nurkholis, [2]Meila Zuwari Negara, [3]Guruh Fajar Shidik, [4]A. Zainul Fanani, [5]Muljono, [6]Edi Noersasongko
Faculty of Computer Science
University of Dian Nuswantoro
Semarang, Indonesia
[1]nurkholis.lldikti6@ristekdikti.go.id, [2]meilazuwari@gmail.com, [3]guruh.fajar@research.dinus.ac.id,
[4]a.zainul.fanani@dsn.dinus.ac.id, [5]muljono@dsn.dinus.ac.id, [6]edi-nur@dsn.dinus.ac.id

***Abstract*— A news sentence is a sentence that tells or does something. In fact, The comparison between positive news and negative news broadcast on TV stations in Indonesia is 1:11. One point for positive newscasting and 11 points for negative newscasting. In the TVKU dataset because the original record number is too large, then the dataset used in this study only used 2.048 data. Which 1025 Positif and 1023 negative. Then obtained accurate results between labeling and testing. By using deep learning, we can get accurate results to analyze the sentiment from TVKU news.**

Keywords -- Deep Learning, MLP

I. INTRODUCTION

The news industry is one of the fastest growing technology industries. The growth of telecommunication services in Indonesia is one of the highest services in the national economy compared to other sectors. Communication is one of the factors used in calculating inflation. The development of the news industry is marked by the increasing number of business people and the increase in customers of telecommunications services.

A news sentence is a sentence that tells or does something. Sentence news can be divided into 2 (two) things, namely positive news sentences and negative news sentences. The positive news sentence is a sentence in which there is no denial. While the negative news sentence is the existence of denial. The comparison between positive news and negative news broadcast on TV stations in Indonesia is 1:11. One point for positive newscasting and 11 points for negative newscasting.

Sentiment Analysis is one of the methods used in information need in social media against the subject [9]. There are many ways of doing sentiment analysis, and machine learning is more popular for this because the approach is based on word semantics, using the list of positive words as well as a list of negative words that have been assigned to be extracted into new semantics. The reason we present this paper it to examine and analyze news systems at TVKU, because it is one of information media which present Indonesian News. It is hoped that this can be taken into consideration for further development of TVKU. To see how many comparisons of positive and negative news on TVKU, the method we proposed based on previous research, we propose sentiment analysis of Indonesian news using deep learning.

II. LITERATURE REVIEW

A. Related Findings

To do classification accuracy of sentiment analysis, many researchers have conducted research. Deep learning is very important in supervised learning and unsupervised learning. Research by B. Swathi Lakshmi, P. Here Raj and R. Raj Vikram [1], proposed unsupervised learning combined with CNN to perform sentiment analysis. Obtained accuracy and accuracy seen increased. It appears that CNN-KNN works better for a smaller dataset and for larger ones it is recommended to use CNN-K Means.

While other studies [2] were conducted using a dataset of over 156,000 phrases described in 1 of about 8,500 sentences. Each phrase has a label-related sentiment. It was concluded that Multinomial Naïve Bayes was better for obtaining basic classification results.

In the study [3] the results obtained that the Deep learning method is not necessarily better than the traditional method of machine learning. Because their performance depends not only on the unity of the network, but also on the size of the training dataset. Accuracy of classification in real settings is much lower than the experimental setting. The SVM method is better than NLP if dataset size is limited.

In research [4], the study attempted long-term neural networks to determine the value of tweet polarity on a global level. And the results show that pretraining unsupervised improves the performance of deep learning.

Meanwhile, according to Deepanway Ghosal et al in his research [5], proposed an ensemble-based model that incorporates deep learning such as CNN and feature-based LSTM to identify optimists and pessimists regarding firms and financial stocks. By combining predictions from various models using ANN to determine opinions. The results of this study were obtained in the form of equations and ranked 2 and 7 for the best team.

B. Sentiment Analysis

Sentiment Analysis is a computational study done by processing the word (text mining) concerned the problem about identifying polarities in user-generated content [6]. Increasing the mass media platform, the more data will be generated.

In general, sentiment analysis can also be said to be a form of classification of text documents to various groups. Often, we only need to classify documents into positive and negative classes [9].

C. Text Mining

Text mining is a process of textual data for extracting and searching information. This area of research is interesting because it is a process of finding information from unstructured text [13]. Text mining makes an article known based on its type based on the words contained in the text. The word is then analyzed and matched on a predefined keyword. Then can be grouping a document efficiently and briefly.

D. Basic Dictionary and Indonesian Stopword List

For those who are in Information Retrieval, information usually requires root words and stop lists (or stop lists). The basic words of all must have been known, namely words that have not changed the word (morphological) such as adding prefixes, suffixes, infixes, inflections, reduplications, etc. Glossary is a general word that does not provide important information (which is usually ignored or discarded, for example, in the indexing process) [15].

E. Text Preprocessing

Text Preprocessing is a process to changing data. Text preprocessing has an important position in text mining [13]. The preprocessing process is a stage to stop the elimination of words, stemming and TF-IDF algorithm. The results of preprocessing text must be numeric.

F. TF-IDF

TF-IDF (is a text mining technique used to group how many words exist in a document by assigning weight to each word, to calculate the weights, we first calculate each word TF with the weight of each word [14].

Term Frequency is a term for terms t and document d determined by: [16]

$$tf_{d,t} = \frac{n_{d,t}}{|d|} \quad (1)$$

where nd, t represents the result of the number of occurrences of the term t in the document d, while the frequency requirement tfd, t represents the result of the number of occurrences of the term t in the document d which is then divided by the number of tokens contained in the document [14]. The reverse document frequency of the term t in the entire the collection[14] ,

$$idf_t = \log \frac{td}{df} \quad (2)$$

IDF (t) is the IDF value of each the word you want to search, but the number the entire document, df number words in all documents [14].

G. Deep Learning

This is a method that uses a set of algorithms to model high-level abstractions in data using complex architectural structure models consisting of several non-linear transformations [11].

Traditional machine learning algorithms can be designed using in-depth learning, but not necessarily the opposite. Neural networks can capture data with very complex characteristics without the hard involvement of manual work. Deep learning uses a deep neural network to learn good representation of data input.

Multi Layer Perceptron is a feed-forward layered network of artificial neurons, where the data circulates in one way, from the input layer to the output layer [21], shown in the following figure.

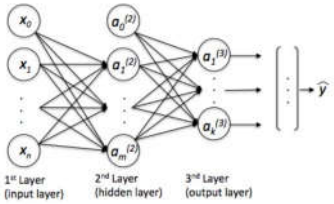


Fig.1
Multilayer Preceptron

III. Proposed Method

A. Dataset

The dataset in this study is a news title that has been aired on TVKU which has been broadcast on April 26, 2018 at 3:59 P.M. until 22 April 2019 at 17.33.33. Because the original record number is too large, then the dataset used in this study only used 2.048 data. The data consists of 1025 positive news titles and 1023 negative news titles. This amount has exceeded the minimum limit for the ideal portion of a study.

B. Method that will be used

The method used in this study is to use deep learning, Deep Learning is based on artificial neural network multilayer feeds trained with stochastic gradients derived using back propagation. With the following stages as follows.

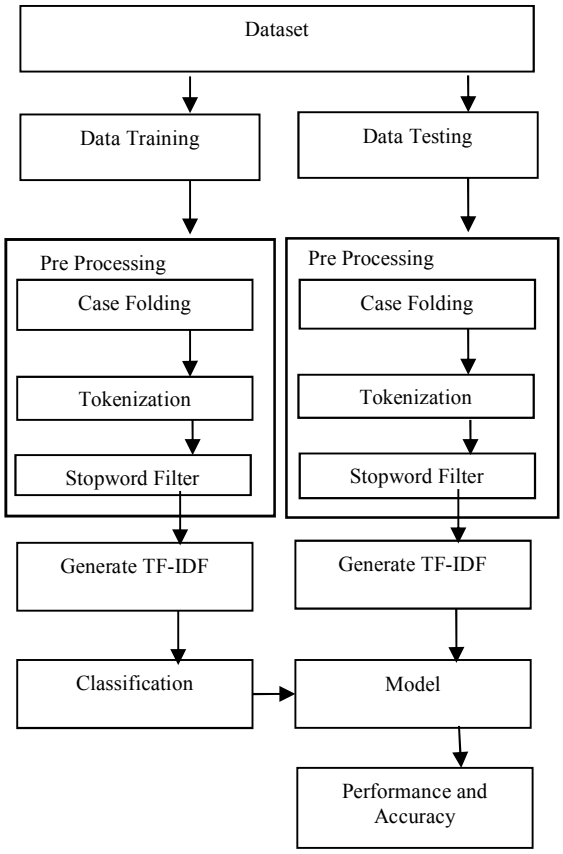


Fig. 2
Model Method

In the method proposed in this study, the cross validation method is carried out where to evaluate the dataset algorithm is separated into 2 ways, namely training data (learning process data) and data testing (validation / evaluation data) to achieve the desired predictions. Data testing will be a model for deep learning. Meanwhile, training data serves to test the performance and truth in this deep learning model.

The next step is to do the preprocessing stage. Transforming input data into results that are ready to be processed. The better the quality of the data to be processed, the better the performance of the classification algorithm. The next stage is tokenisasi, in this process the solution to the headline is done to become a unit of words. At this stage, every space in the news title is based on the space that the word can be broken down.

Stopword filter is a process to remove unnecessary words. For example, in the news title that will be processed will be removed the word "in", "to". With stop word, it will help to reduce the number of words that must be processed. Then the next step is generating TF-IDF. This is done to find the word in the headline where it is heavier.

IV. Experiment and Discussion

In doing this research used computer specifications with Intel® Core ™ i7-4510u processor, 2GHz, 8 GB memory, 500 GB SSD hard drive, Windows10 operating system and RapidMiner application ver 7.6.001.

The dataset consists of 2048 TVKU news titles. Examples are as follows:

TABLE I
DATASET NEWS TITLE

No	Title
1	Relokasi PKL BKT, 2. 700 Bangunan Liar Telah Dirobohkan
2	Kerajinan Kayu dari Limbah Kayu Jati
3	Peringati Haul, Pemkab Blora Bersih-bersih Kediaman Pramudya
4	Lomba Burung Berhadiah Emas
5	Polres Jepara Gerebeg Gudang Miras

Then do it by labeling the data (news title), given a "positive" or "negative" label. The next step, the preprocessing stage is carried out. Beginning with case folding to convert the text in a news title into a standard form. The examples are as follows:

TABLE II
Case Folding Result

No	Title
1	relokasi pkl bkt bangunan liar telah dirobohkan
2	kerajinan kayu dari limbah kayu jati
3	peringati haul pemkab blora bersih bersih kediaman pramudya
4	lomba burung berhadiah emas
5	polres jepara gerebeg gudang miras

Next, tokenizing, this processes each news title into words per word. Like the following table:

TABLE III
Tokenizing Result

No	1	2	3	4	5
T1	relokasi	kerajinan	peringati	lomba	polres
T2	pkl	kayu	haul	burung	jepara
T3	bangunan	dari	pemkab	berhadiah	gerebeg
T4	liar	limbah	blora	emas	gudang
T5	telah	kayu	bersih		miras
T6	dirobohkan	jati	bersih		
T7			kediaman		
T8			pramudya		

Stopword filtering to get important words in the news title to be able to proceed in the next process. Table III is filtered so as to get the following results:

TABLE IV
Stopword Filtering Result

No	1	2	3	4	5
T1	relokasi	kerajinan	peringati	lomba	polres
T2	pkl	kayu	haul	burung	jepara
T3	bangunan		pemkab	berhadiah	gerebeg
T4	liar	limbah	blora	emas	gudang
T5		kayu	bersih		miras
T6	dirobohkan	jati	bersih		
T7			kediaman		
T8			pramudya		

Then continue generating TF-IDF, each news title is calculated by its weight value.

We use 3 hidden layer which 50, 200, 50 neurons. Accuracy calculations are performed by calculating true positive and true negative values, based on positive predictions and negative predictions.

In this research, we use the operators which available on RapidMiner, as shown in the following picture.

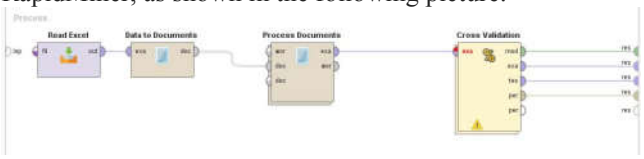


Fig 3.
Sentiment Analysis Process with Rapid Miner,

The details of the cross validation process are as shown below.

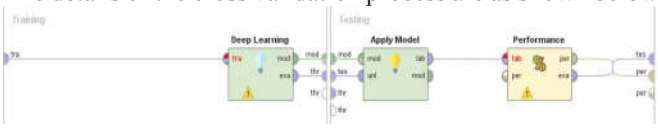


Fig.4
Cross Validation Process

Based on example set (2048 examples, 4 special attributes, 4013 regular attributes), cross validation data was obtained. From 2048 example, we found 471 wrong predictions. Shown in the following picture.

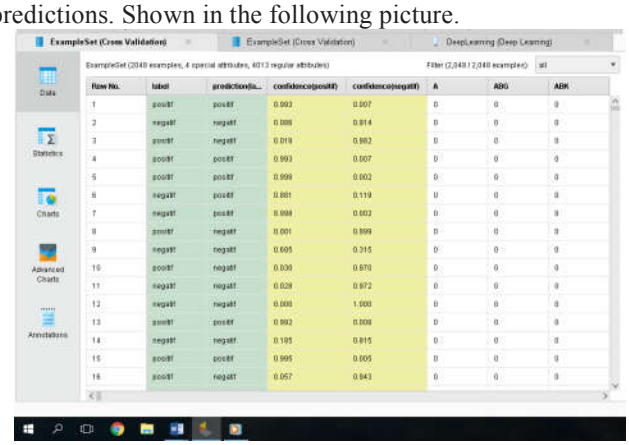


Fig.5
ExampleSet (Cross Validation)

The statistic result show from 2048 data, there is 1025 positive data were labeled and 1023 negative data were labeled. After cross validation testing, it was found that the prediction (label) which positive news was 1144 data and negative news was 1144 data. Shown in the following table.

TABLE V
CROSS VALIDATION STATISTIC

Name	Least	Most	Value
Label	Negative (1023)	Positive (1025)	Positive (1025), negative (1023)
Prediction	Negative (904)	Positive (1114)	Positive (1144), negative (904)
Confidence (positive)	Min 0.0000	Max 1.0000	0.529
Confidence (negative)	Min 0.0000	1.0000	0.471

In the following figure, the ROV curve is shown, a plot formed from the true positive rate. AUC: 0.859 +/- 0.013 (micro average: 0.858) (positive class: negative) closer to the value 1 indicating a predictive match.

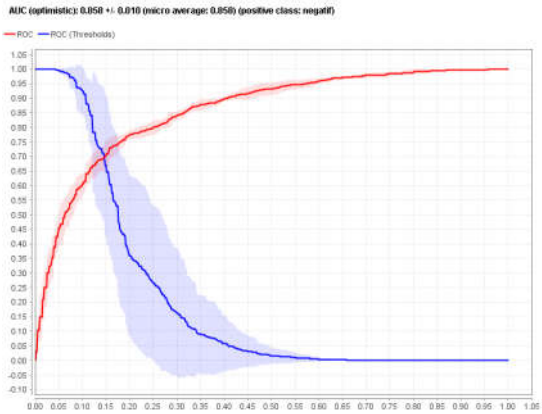


Fig.6
ROV Curve

In the following table shows the predicted values match the actual value when testing. Based on predictions, where prediction positif, the result show that true positive is 814 and 244 is true negative. And the matrix shows that the accuracy of the model is for a positive prediction of 76.94% and the accuracy of the negative prediction model is 78.69%. The prediction accuracy show 77.00% +/- 2.09% .

TABLE VI
PERFORMANCE VECTOR

	True positif	True negative	Class precicion
Pred positif	849	295	74.21%
Pred negative	176	728	80.53%
Class recall	82.83%	71.16%	

V. CONCLUSIONS

A news sentence is a sentence that tells or does something to others. Word can be divided into 2 (two) things, namely positive news sentences and negative news sentences. The comparison between positive news and negative news broadcast on TV stations in Indonesia is 1:11. One point for positive newscasting and 11 points for negative newscasting. The data set used in this study is a collection of data from TVKU. So the dataset used in this study uses only 2,048 data. By using in-depth learning, we will get the right value to analyze TVKU news. Obtained an accuracy of 77.00%.

REFERENCES

[1] B. Swathi Lakshmi, P.Sini Raj dan R.Raj Vikra, "Sentiment Analysis Using Deep Learning Technique CNN with KMeans", International Journal of Pure and Applied Mathematics, Volume 114, No. 11 2, page 47-57, 2017.

[2] Qurat Tul Ain, Mubashir Ali , Amna Riaz , Amna Noureen , Muhammad Kamran , Babar Hayat dan A. Rehman., "Sentiment Analysis Using Deep Learning Techniques: A Review", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 6, 2017

[3] Yujie Lu, Kotaro Sakamoto, Hideyuki dan Shibuki Tatsunori Mori., ″Are Deep Learning Methods Better for Twitter Sentiment Analysis? The Association for Natural Language Processing", 言語処理学会 第23回年次大会 発表論文集 (2017年3月)

[4] Miguel A. Alonso , Yerai Doval, David Vilares, and Carlos G´omez-Rodr´ıguez, ″Deep Learning Experiments for Sentiment Analysis on Spanish Tweets″, TASS, pp 47-52, 2015.

[5] Asif Ekbal, Pushpak Bhattacharyya, "Task 5: An Ensemble of Deep Learning and Feature Based Models for Financial Sentiment Analysis″, Proceedings of the 11th International Workshop on Semantic Evaluationss, pages 899–903, 2017.

[6] Jan Deriu, "Sentiment Analysis using Deep Convolutional Neural Networks with Distant Supervision″, Thesis, Department of Computer Science, ETH Zurich, 2016.

[7] Md Shad Akhtar , Ayush Kumar, Asif Ekbal, Pushpak Bhattacharyya, "A Hybrid Deep Learning Architecture for Sentiment Analysis. Proceedings of COLING 2016", the 26th International Conference on Computational Linguistics: Technical Papers, pages 482–493, 2016,

[8] Mihaela Sorostinean, Katia Sana, mohamed Mohamed, Amal Targhi, "Sentiment Analysis on Movie Reviews″, Journal Agroparistech 2017.

[9] Sahar Sohangir, Dingding Wang, Anna Pomeranets and Taghi M. Khoshgoftaar., "Big Data: Deep Learning for financial sentiment analysis″ Journal of Big Data *5:3,* 2018.

[10] O` scar Romero Llombart, "Using Machine Learning Techniques for Sentiment Analysis", Final Project On Computer Engineering, School Of Engineeriing(EE), Universitat Auto` Noma De Barcelona (UAB), 2017.

[11] Prerana Singhal, Pushpak Bhattacharyya, "Sentiment Analysis and Deep Learning: A Survey″, Indian Institute of Technology Bombay, 2017.

[12] Houshmand Shirani-Mehr, " Applications of Deep Learning to Sentiment Analysis of Movie Reviews″, CS224d: Deep Learning for Natural Language Processing. Stanforf University, 2015, Access from https://cs224d.stanford.edu/reports/Shirani-MehrH.pdf.

[13] Dr. S. Vijayarani, Ms. J. Ilamathi, Ms. Nithya, "Preprocessing Techniques for Text Mining - An Overview", International Journal of Computer Science & Communication Networks, Vol 5(1), pg 7-16, 2015.

[14] Bening Herwijayanti, Dian Eka Ratnawati, Lailil Muflikhah, "Klasifikasi Berita Online dengan menggunakan Pembobotan TF-IDF dan Cosine Similarity″, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, Vol. 2, No. 1, Januari 2018, page 306-312, 2018.

[15] Putra Pandu Adikara, "Kamus Kata Dasar dan Stopword List Bahasa Indonesia″, Access from http://hikaruyuuki.lecture.ub.ac.id/kamus-kata-dasar-dan-stopword-list-bahasa-indonesia/

[16] Nils Schaetti. "UniNE at CLEF: TF-IDF and Deep-Learning for Author Profiling Notebook for PAN″ *CLEF 2017*, access from ceur-ws.org/Vol-1866.

[17] tvku.tv/ date

[18] Djoko Budiyanto Setyohadi, Setyohadi, Felix Ade Kristiawan, Ernawati, "Perbaikan Performansi Klasifikasi dengan Preprocessing Iterative Partitioning Filter Algorithm″, Telematika, 2017, Vol 14, No.01, pp 12-20, 2017

[19] Ira Zulda, Edi Winarko, "Sentimen Analisis Tweet Berbahasa Indonesia dengan Deep Belief Network", IJCCS, 2017, Vol 11 No.2. July 2017. pp 187-198.

[20] Hirroe Wijaya Ani Kesuma, "Penerapan Metode TF-IDF dan Cosine Similarity dalam Aplikasi Kitab Undang-Undang Hukum Dagang″, Unnes Semarang, 2017.

[21] Syrine Ben Driss, "A comparison study between MLP and Convolutional Neural Network models for character recognition", 2017, access from https://hal-upec-upem.archives-ouvertes.fr/hal-01525504/document

Measurement of Spinal Curvature for Scoliosis Classification

Sinta Kusuma Wardani [1)], Riyanto Sigit [2)], Setiawardhana [3)], Seffiana Manik Syah Putri [4)], DindaAyu Yunitasari [5)]
Informatic and Computer Departement
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
sintakusuma@ce.student.pens.ac.id [1)], riyanto@pens.ac.id [2)], setia@pens.ac.id [3)], seffiana.putri8@gmail.com [4)],
dindaayuyunitasari@ce.student.pens.ac.id [5)]

Abstract— Scoliosis is a spine disorder phenomenon that makes spine bend and it will form letter C or S. Manual method in determining the curvature of the spine results in a very low accuracy value. This occurs because of the noise present in the spinal x-ray images of patients with scoliosis. The noise can be either organ, blood, or bone that makes the calculation of curvature of the spine to be inaccurate. Therefore in this research the application of scoliosis classification based on the level of curvature of the spine will be made using images processing approach. The spinal curvature classification consists of several processes. The process begins by preprocessing the spinal bone image using the median filter and morphological watershed method. Then it proceeds with features extraction and then the process of classification begins with artificial neural network method. It is expected that the classification of curvature of the spine to be more accurate and can be handled in accordance with its classification. From the experimental system found 0.26 % error for spinal segmentation using watershed.

Keywords— spine; image processing; scoliosis; watershed; artificial neural network

I. INTRODUCTION

Scoliosis is an abnormally curved condition of the spine. About 2-4% of the adolescent population has some degree of scoliosis. About 2.2% of these adolescents will require treatment, which consists of observation, orthotic (clamp), or surgical treatment [6].

Treatment decisions for scoliosis patients are based on consideration of the degree of spinal curvature [2]. Currently the process of determining the angle of the spine is still often done manually by orthopedic doctors. The thing that often happens is the measurement results between doctors with one another can be different because there is still a lack of accuracy in the process of determining cobb angle manually[3].

Therefore, in this research will be classified scoliosis based on curvature of the spine. By doing the classification of scoliosis based on bone curvature it will facilitate the doctor in determining treatment decisions to be performed on the patient.

Segmentation is done by using watershed method with x-ray of spinal image of scoliosis patient. Segmentation aims to get a curved spinal image object [7]. After that calculation of curvature of spinal image by using Euclidean Distance method

Artificial Neural Network is one of the methods used in classification[5]. In this study the artificial neural network method is used to classify scoliosis based on the curvature of the spine. This method shows the result of classification according to the desired. There are two stages in the classification by using ANN, are the training phase and the testing stage. At the testing stage, the pattern as input on the ANN produces output which presents what is believed by ANN as the correct output. Nevertheless, for a successful test phase should be preceded by a training phase. At this stage the ANN receives a set of entries corresponding to a set of outputs. In other words, ANN learns to pair the correct outputs to the corresponding input patterns [2].

II. METHODOLOGY

To measure the distance of the curvature of the spine, the procedure is load the image, do preprocessing, segmentation and then calculate the distance curvature. A procedure to segmentation and measure the curvature of the spine in this system described by figure 1.

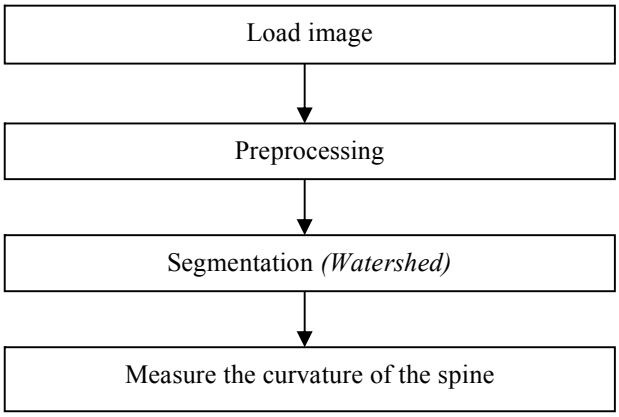


Fig. 1. System Diagram

A. Load Image

In this research the first process is to read the input data from the spinal x-ray image of the scoliosis patient. Figure 2 is a x-ray image of scoliosis patient.



Fig. 2. X-Ray Image

.

B. Preprocessing

Preprocessing is process to enhance and improve the image quality of radiology image that have a lot of noise. These are some steps for *preprocessing*:

1) Histogram Equalization

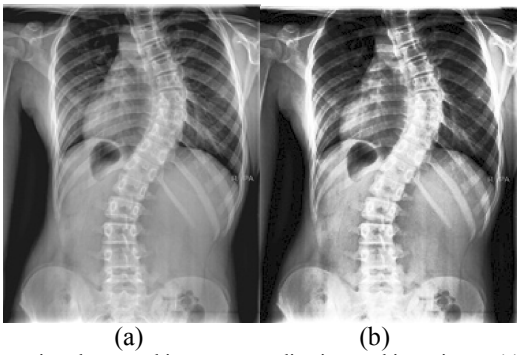
Histogram equalization is a histogram alignment process, in which the gray-degree value distribution in an image is made flat. Histogram equalization also a process where the histogram is leveled based on a linear function (straight line). To be able to do this histogram equalization, needed a distribution function which is overlap from the histogram [4].

The purpose of histogram equalization is to get a histogram equalization so that each degree of gray has the same relative number of pixels. The histogram equalization value is obtained by changing the gray level of a pixel with a new gray degree with the transformation function. The formula of histogram equalization can be represented as:

.

$$P_r(r_k) = \frac{n_k}{n}, \text{ when } r_k = \frac{k}{L-1}, 0 \leq k \leq L-1 \quad (2)$$

result of histogram equalization process can be seen in a Figure 3



(a) (b)

Fig. 3. Comparison between histogram equalization and input image (a) Input image (b) Histogram equalization

2) Median Filtering

After going through the equalization process, the next image is processed by using median filter. Median filtering is a non-linear filter commonly used to reduce noise and can maintain the edge of the image. The median filtering obtained from the original image pixel value is composed from small to large and replaced by the middle pixel value. [12]. The following is the median filter equation:

$$f(x,y) = med\ (p,t) \in Pxy\ \{f(p,t)\} \quad (1)$$

Figure 4 is the result of median filtering process



(a) (b)

Fig. 4. Comparing median filter with input image(a) Input image (b) Median filter

C. Segmentation

Watershed method is the method used to do the segmentation, so that only produce spinal image only. The watershed process begins with the drawing or initialization of watershed points. This point will determine the segmentation area of the image. The system will determine 2 segmentation areas[9]. Area consists of:

1. Area inside the spine
2. Area outside the spine

Image intensity is used as a reference in a watershed algorithm. This intensity will be used to convert images into topography of valleys and mountains. The valley part is a value that indicates low image intensity, while the mountain part indicates that the image intensity is high.

Markers are used to mark the topographical area that will be flooded, as if the marker is the source of the springs of the topography. The water source will flood the valley on the topographical area based on the marker made. Water flows from the lowest point in the topographic area, the water flows to the same height as the other markers. After reaching the same height fill the water together until you see the peak [1]. The result of watershed segmentation represented by figure 5

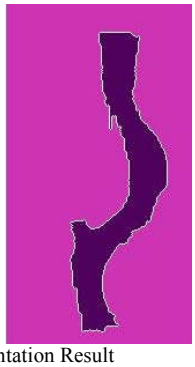


Fig. 5. Watershed Segmentation Result

D. Crop

Cropping process serves to remove the remaining noise after segmentation process. The cropping process is taken from watershed segmentation. The watershed output is used as a masking to cut the area in the original image. So obtained the results of cutting the area of the spine. The result of the watershed segmentation of the previous process in convert first becomes gray scale. Then the image will be processed with threshold with binary threshold type. The system will then make the background white to look contrast with the crop object. The image to be cropped and past thresholding is then copied to the prepared background. The result of cropping represented by figure 6.



Fig. 6. Crop Operations Result

E. Threshold Binary

Thresholding applied to this system is Binary Thresholding, which is Thresholding with two colors (black and white). The function of thresholding is used to identify and separate the desired object from the background image based on gray level distribution or image texture [11]. The algorithm for binary thresholds is that if the pixel value is smaller than the threshold, the pixel value will be replaced by 0 and if the pixel value in the image is greater than the threshold value, the pixel value will be replaced with the maximum value. Threshold is used to segment the entire spine image [8].The result of threshold binary represented by figure 7.

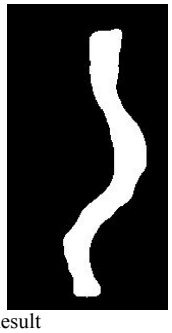


Fig. 7. Threshold Binary Result

F. Thinning

One of the thinning algorithms is Zhang Suen. This algorithm is a simple thinning algorithm. First, below is the condition given by Zhang Suen's algorithm:

1. 2 <= N (G1) <= 6
2. T (G1) = 1
3. G2 * G4 * G6 = 0
4. G4 * G6 * G8 = 0
5. 2 <= N (G1) <= 6
6. T (G1) = 1
7. G2 * G4 * G8 = 0
8. G2 * G6 * G8 = 0

The above has presented the conditions given by Zhang Suen's algorithm. Next make a rule, so this rule should be followed. We will use the binary image, where the white pixel as the foreground and the black pixel as the background [10].

G. Calculate The Curvature of The Spinal Image

To calculate the curvature of the spine is used distance and standard deviation parameters. First to find the distance used euclidean distance equations and second to find standard deviation used find contour command to find the position value of x from the image then the value is sought standard deviation.

$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{(n-1)}} \quad (3)$$

Where,
s = standard deviation
n = number of samples

$$|Dist| = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (4)$$

Where
x1 = x maximum coordinates
x1 = x minimum coordinates
y1 = y coordinates (0)
y1 = y coordinates (0)

III. Experimental And Result

This section will be explain about the final results of the system. The data used as input for this research were x-ray data of scoliosis patients obtained from the radiology unit of Surabaya Haji Hospital. The results of this study are explained

in table form. In Table I is the result of the segmentation process with the watershed method and image cropping. In Table II is a comparison of segmentation results using watershed algorithm and find contour. Next in Table III is a comparison between the results made by the system and manual calculations. From this table shows that the accuracy generated by very high systems is almost close to manual calculation.

TABLE I. WATERSHED SEGMENTATION AND CROPPING

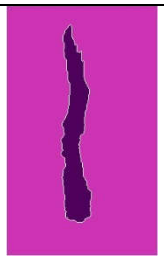
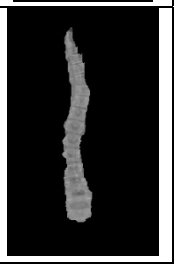
Input	Watershed	Crop
		
		
		

TABLE II. COMPARISON BETWEEN FIND CONTOUR AND WATERSHED METHOD

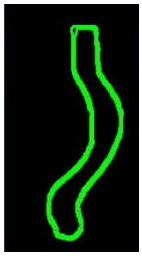
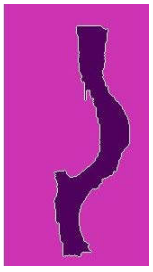
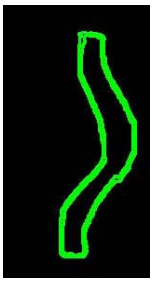
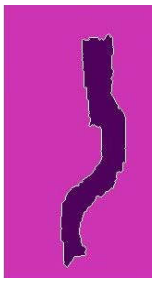
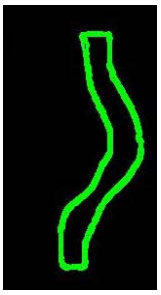
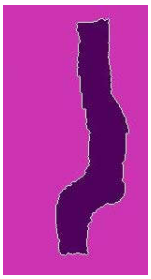
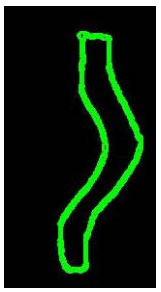
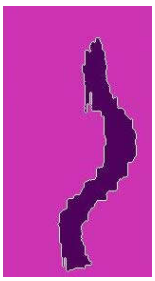
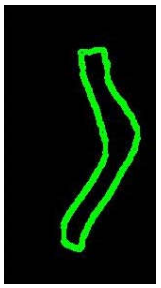
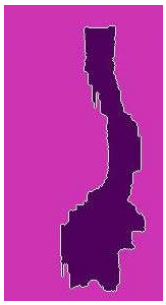
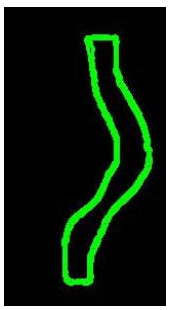
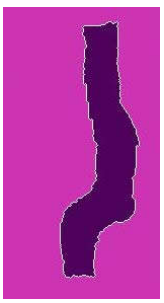
Picture	Filter (kernel)	*Find Contour*	*Watershed*
1.	Median (27)		
2.	Median (37)		
3.	Median (47)		
4.	Gaussian (17x17)		
5.	Gaussian (27x27)		
6.	Median (47)		

TABLE III. DISTANCE CALCULATION

Data	Manually	Watershed	Error
1	5	5.2	0.2
2	4	4.3	0.3
3	3	3.1	0.1
4	4	4.5	0.5
5	10	10.2	0.2
Mean			0.26

IV. CONCLUSION AND FUTURE WORK

The purpose of making this system is for segmentation the spinal image then calculate the curvature of the spinal image to knowing scoliosis classification. Preprocessing is use to enhancement and improve the image quality using Median Filtering and Histogram Equalization. The segmentation using Wateshed. Then to calculate the spinal curvature using Euclidean Distance. This system has a 99.74% presentation of success. For future research, can do optimization by automatically segmentation using watershed algorithms to improve the accuracy of results.

ACKNOWLEDGMENT

The authors thank the Politeknik Elektonika Negeri Surabaya for facilitating this research and also thank the Surabaya Haji Hospital for providing medical information, data and discussions. In addition, thanks to Mr. Riyanto Sigit and Mr. Setiawardhana for the discussions that were useful and very helpful for this research.

REFERENCE

[1] Rizal Romadhoni Hidayatullah, Riyanto Sigit, Sigit Wasista, "Segmentation of Head CT-Scan to Calculate Percentage of Brain Hemorrhage Volumes,"2017, International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), Surabaya, 2017.

[2] C.SATHYA, B.SOWNDARYA, "Measurment of Spine Curvature for The Detection and Classification of Scoliosis Using Matlab," 2016 International Journal of Advanced Research Trends in Engineering and Technology (IJARTET) Vol. 3

[3] Cheung C. J, Zhou G, Law S, Mak T, Lai K, and Zheng Y., "Ultrasound Volume Projection Imaging for Assessment of Scoliosis," 2015, IEEE transactions on Medical Imaging. TMI-2014-1053.

[4] Joseph A Janicki, Benjamin Alman. 2007. *Scoliosis: Review of Diagnosis and Treatment*. Paediatr Child Health. Nov 12(9): 771–776.

[5] Junhua Zhang.Edmond Larence.dkk.2009. "Automatic Automatic Cob Measurment of Scoliosis Based on Fuzzy Hough Transform with Vertebral Shape Prior". Journal of Digital Imaging,Vol 22, No 5

[6] Kevin Law.2016. "Rungkup dan Pembedahan Tulang Belakang Tak Efektif dalam Perawatan Skoliosis".

[7] Mujahidin, Alif Aziz, 2014, "Segmentasi Citra Tulang Belakang Menggunakan Charged Particle Model"

[8] Negrini S, De Mauroy JC, Grivas TB, Knott P, Kotwicki T, Maruyama T, et al. "Actual Evidence In The Medical Approach To Adolescents With Idiopathic Scoliosis". *Eur J Phys Rehabil Med*. 2014;50(1):87–92.

[9] Putri Famela Azhari. Bambang Hidayat. Achmad Riza P. 2013. "Simulasi dan Analisis Perhitungan Derajat Kelengkungan Tulang Punggung Pada Manusia Menggunakan Metode Trasformasi Contourlet dan K-Nearest Neighbor". Teknik Telekomunikasi, Fakultas Teknik Elektro, Universitas Telkom

[10] Raka Kundu. Prasanna Lenka. 2012. "Cobb Angle Quantification for Scoliosis Using ImageProcessing Techniques. International Conference on Recent Advances and Future Trends in Information Technology" *(iRAFIT2012)*

[11] Susilo, Sunarno, I Ketut Swakarma, dkk. 2013. "Kajian Sistem Radiografi Digital Sebagai Pengganti Sistem Computes Radiography yang Mahal". Jurnal Fisika Indonesia No:50, vol.XVII, ISSN:1410-2994.

[12] Tariq Abuzaqhleh dan Buket Barkana. 2012. "Computer Aided Technique for The Measurment of The Cobb Angle"

Mining Medication Behavior of The Completion Leprosy's Multi-Drug Therapy in Indonesia

[1,5*] Enny Rachmani, [1,2**]Ming-Chin Lin, [3,4***] Chien-Yeh Hsu, [6****]Guruh Fajar Shidik, [6*****]Edi Noersasongko

[1]Graduate Institute of Biomedical Informatics Taipei Medical University, Taipei, Taiwan

[2]Division of Neurosurgery, Department of Surgery, Taipei Medical University-Shuang Ho Hospital, Zhonghe District, New Taipei City

[3]Department of Information Management, National Taipei University of Nursing and Health Science, No.365, Ming-te Road, Beitou District, Taipei City, Taiwan

[4]Master Program in Global Health and Development, Taipei Medical University. Xinyi District, Taipei City 101, Taiwan

[5]Department Health Information Management, Faculty of Health Science, Universitas Dian Nuswantoro. Semarang

[6]Department Informatics, Faculty of Computer Science, Universitas Dian Nuswantoro, Semarang, Indonesia

Email: *enny.rachmani@dsn.dinus.ac.id; **arbiter@tmu.edu.tw; ***cyhsu@ntunhs.edu.tw; ****guruh.fajar@research.dinus.ac.id; *****edi-nur@dsn.dinus.ac.id

Abstract— Leprosy is one of the listed neglected tropical diseases which is continuing as a major health problem in Indonesia. Low adherence to Multi-Drug Therapy (MDT) against leprosy is still a significant obstacle in disease control, as the defaulters with incomplete cure remain as potential sources of infection, lead to patients' irreversible complications and multidrug resistance.

This paper aims to describe the behavior of leprosy patients during the long-term medication of MDT drugs through mining association rule between the on-time of taking MDT to healthcare facilities using apriori algorithm. Data has been collected from manual leprosy registry cohort of 27 Primary Health Care (PHC) at Pekalongan District start from June, 1st 2012 to June, 30th 2014.

This study converted the range date between the MDT collection attendances to the new label, which is called on-time attendance (OTA), and on time completion (OTC). Both of MB and PB patients shows the association between MDT 2, MDT 3 have a strong relation to the late of the completion treatment behavior of leprosy patients.

Keywords—Association rule; Apriori; Multi-Drug Therapy; Leprosy; Primary Health Care.

I. INTRODUCTION

Leprosy is one of the listed neglected tropical diseases which is continuing as a major health problem in Indonesia. Of the 200,000 new cases of leprosy occurring worldwide each year, approximately 80% occur in India, Brazil, and Indonesia. Indonesia reported a prevalence of 20,160 cases, with the number of new cases in 2015 at 17,202 [1]. The fact that only 45% of leprosy patients attended clinic regularly, 22.5% are irregular, and 32.4% have discontinued the treatment [2].

WHO recommended Multi-Drug Therapy (MDT) for treating leprosy patients. Leprosy patients should complete MDT consisting of rifampicin, clofazimine, and dapsone for a 6- to 12-month period, depending on the leprosy type. Leprosy Paucibacillary (PB) type should consume six MDT blister packs while Multibacillary (MB) type should consume for 12 packs of MDT. Each MDT blister pack contains pills for four weeks consumed [3]. Low adherence to MDT against leprosy is still a significant obstacle in disease control, as the defaulters with incomplete cure remain as potential sources of infection, lead to patients' irreversible complications and multidrug resistance [4-6].

Clinical computing has aims to provide decision support by capturing and analyzing clinical data to increase medical knowledge. The identifying process to find novelty and potentially patterns in data which meaningful has been described as "data mining" and "knowledge discovery" [7, 8].

This paper aims to describe the behavior of leprosy patients during the long-term medication of MDT drugs through mining association rule between the on-time of taking MDT to healthcare facilities using apriori algorithm. Association rules algorithm has often been used to explore medical knowledge in health area [8-10]. Understanding a patient's behavior could improve the quality of medical treatment especially for the prolonged treatment of leprosy patients by preventing irregular treatment and the drop out medication.

II. METHOD

A. Data Collection

Primary data has been collected from manual leprosy registry cohort of 27 Primary Health Care (PHC) at Pekalongan District start from June, 1st 2012 to June, 30th 2014. Data has 124 patients who completed the treatments during the period. Each patient has 6 or 12 attributes of MDT collection. Each PHC has leprosy cohort patients which manually recorded and keep in the PHC.

In leprosy cohort data, each patient recorded individually during the treatment 6 or 12 months consistent with their leprosy type and the monitoring data of for the next five years.

B. Pre-processing

Manually leprosy registry cohort in PHC consists of the attendances date of leprosy patients every month during the treatment consistent with the leprosy type. Leprosy disease has two types consist of leprosy PB patients who have sixth-month date attendances while MB type has 12 date attendances.

This study has converted the range date between the MDT collection attendances to the new label, which is called on-time attendance (OTA), and on time completion (OTC). The OTA label has been given if the range date between MDT <31 days while OTC label has given to patients whose collected MDT in six or 12 months according to the leprosy type. The patient collected more than the period mention above classify as late.

C. Item Set

Each patient has items set of the on-time attendances during taking MDT drug from PHC, shown in Table 1. It consists of MDT one to MDT six or MDT twelve consistent with their leprosy type. Patients also have attributed the completion of the MDT collection (Release from Treatment=RFT).

Table 1. The features and label of the data

Item	On Time	Late
MDT 01	MDT01OTA	MDT01LA
MDT 02	MDT02OTA	MDT02LA
MDT 03	MDT03OTA	MDT03LA
MDT 04	MDT04OTA	MDT04LA
MDT 05	MDT05OTA	MDT05LA
MDT 06	MDT06OTA	MDT06LA
MDT 07	MDT07OTA	MDT07LA
MDT 08	MDT008TA	MDT08LA
MDT 09	MDT009TA	MDT09LA
MDT 10	MDT010TA	MDT10LA
MDT 11	MDT011TA	MDT11LA
MDT 12	MDT012OTA	MDT12LA
RFT	RFTOTC	RFTLC

D. Association Rules Apriori

Association rule has been used to express the relationship between item sets to each other or known as $X \rightarrow Y$. Both X and Y are item set. Apriori is one of association method, which counts the frequent probability value of relation item set that should meet with minimum the condition [11]. The minimum condition should be meet the required parameters. Support and confidence are the parameters that showed quality the association rules.

Support is a condition when the sum of the relation of each item set could meet with some minimum parameter threshold. It is vital to decide the frequent item set, where the formula could be seen in (1).

$$Support\ (X \rightarrow Y) = P(X \cup Y) \quad (1)$$

The other parameter confidence is essential to decide the quality of strong association rules in each item set. Confidence shows the percentage item set Y in each itemsets X. Detail of parameter confidence could be seen in formula (2).

$$Confidence\ (X \rightarrow Y) = P(X|Y) \quad (2)$$

In order to understand the detail steps of association rule apriori could be seen in [11-13].

E. Tools

The Rapid miner has been used to mining the data set with the w-appriori operation to explore association rule between MDT. During simulation parameter Apriori algorithm is set as:

Table 2. The parameter for mining the association of dataset

Parameters		MB	PB
N	The required number of rules	10	10
T	The metric type by which to rank rules	0.0	0.0
C	The minimum confidence of a rule	0.7	0.5
D	The delta by which the minimum support is decreased in each iteration	0.05	0.05

Parameters		MB	PB
U	Upper bound for minimum support	1.0	1.0
M	The lower bound for the minimum support	0.1	0.1
S	If used, rules are tested for significance at the given level	-1.0	-1.0

III. RESULT AND DISCUSSION

The leprosy dataset has proportion the MB type higher than the PB type. Table 2 shows that MB patient is a higher proportion (65%) than PB type (35%)

Table 2 shows that MDT 3 is a first peak proportion for the late attendance of patients both in MB and PB type. Leprosy patients with MB type have four peaks the late of attendances consist of MDT 3, MDT 6, MDT 8 and MDT 11 while PB type has a peak of late the attendance at MDT 4.

Leprosy patients need to consume drugs for six months or 12 months. The long-term medication period has potential drop out and irregular treatment. Knowing the weak of the long-term medication by identity which MDT has a high contribution for the late of the completion treatment should be acknowledged by health officers to prevent irregular treatment.

Table 3. Distribution frequency of the on-time attendance among MDT

ITEM	On Time		Late	
	f	%	f	%
MB TYPE				
MDT 1	80	100	0	0.0
MDT 2	65	81.2	15	18.8
MDT 3	56	70.0	24	30.0
MDT 4	58	72.5	22	27.5
MDT 5	55	68.7	25	31.3
MDT 6	53	66.3	27	33.7
MDT 7	57	71.3	23	28.8
MDT 8	48	60.0	32	40.0
MDT 9	52	65.0	28	35.0
MDT 10	52	65.0	28	35.0
MDT 11	45	56.3	35	43.7
MDT 12	51	63.8	29	36.3
RFT	41	51.2	39	48.8
PB TYPE				
MDT 1	44	100	0	0.0
MDT 2	30	68.2	14	31.8
MDT 3	24	54.5	20	45.5
MDT 4	22	50.0	22	50.0
MDT 5	25	56.8	19	43.2
MDT 6	30	68.2	14	31.8
RFT	23	52.3	21	47.7

Table 4. Association rule of MB patient treatment

Rules	Rules	Conf
1	MDT 2=Late, MDT 3=Late => RFT= Late	1
2	MDT 4=Late, MDT 5=Late => RFT= Late	0.86
3	MDT 3=Late, MDT 4=Late => RFT= Late	0.78
4	MDT 2=Late => RFT= Late	0.76
5	MDT 3=Late, MDT 6=Late => RFT= Late	0.74
6	MDT 6=Late, MDT 11=Late => RFT= Late	0.74

Table 4 shows that RFT of MB patient has an association with the late of MDT 2 and MDT 3 if patients late in both of them they will develop late in RFT with confidence 1. Patients who're late in MDT 4 and MDT 5 also have a high opportunity late in RFT (C=0.86).

Table 5. Association rule of PB patients

Rules	Rules	Conf
1	MDT 2= Late =>MDT 3=Late	0.56
2	MDT 5=Late => RFT= Late	0.55
3	MDT 4=Late => RFT= Late	0.50

Table 4 and table 5 show the association rule of the on-time completion of the leprosy patients. Both of MB and PB patients shows the association between MDT 2, MDT 3 and developing the late of the completion treatment of leprosy patients. This fact should be a concern for health officers during treatment for leprosy patients. The health officer should encourage leprosy patients especially during MDT 2 and MDT 3 treatments. Irregular treatment of leprosy patients even drop out, or default treatment could develop resistance and increase the disability grade of leprosy patients [14, 15].

IV. CONCLUSION

Indonesia is the third biggest in the world has leprosy patients and always develop thousands of new cases every year. Mining the medication completion of leprosy patients has benefit to understand the behavior of the long-term treatment of leprosy patients especially in Indonesia. The results of an experiment in this work have shown, the on-time completion

medication behavior of leprosy patients of MDT 2 and MDT 3 during treatments have highest confidence relation in both MB and PB cases up to 100%.

The knowledge discovery from this paper should be a concern for leprosy surveillance officers in primary health care to handle the patients and prevent the uncompleted the treatment. The future study should be conducted to mining the related factor of MDT drugs adherences to prevent irregular treatment and developing new cases.

V. REFERENCE

1. Wibawa, T. and T.B.T. Satoto, *Magnitude of Neglected Tropical Diseases in Indonesia at Postmillennium Development Goals Era.* Journal of tropical medicine, 2016. **2016**.
2. Kumar, A., et al., *Utilization of medical agencies and treatment compliance by urban (Madras) leprosy patients.* Leprosy in India, 1983. **55**(2): p. 322-332.
3. Organization, W.H., *Global Leprosy Strategy 2016-2020: Accelerating towards a leprosy-free world*, in *Global Leprosy*. 2016, World Health Organization: Geneva.
4. Rao, P., *A Study on Non-adherence to MDT among Leprosy Patients.* Indian journal of leprosy, 2008. **80**(2): p. 149.
5. SMITH, W.C.S., *Assessing and improving adherence with multidrug therapy.* Lepr Rev, 2012. **83**: p. 282-291.
6. Raju, M., et al., *Correlates of Defaulting from MDT among Leprosy Patients.* Indian J Lepr, 2015. **87**: p. 241-248.
7. Maimon, O. and L. Rokach, *Data mining and knowledge discovery handbook*. Vol. 2. 2005: Springer.
8. Prather, J.C., et al. *Medical data mining: knowledge discovery in a clinical data warehouse*. in *Proceedings of the AMIA Annual Fall Symposium*. 1997. American Medical Informatics Association.
9. Harpaz, R., H. Chase, and C. Friedman, *Mining multi-item drug adverse effect associations in spontaneous reporting systems.* BMC Bioinformatics, 2010. **11**(Suppl 9): p. S7.
10. Obenshain, M.K., *Application of data mining techniques to healthcare data.* Infection Control & Hospital Epidemiology, 2004. **25**(08): p. 690-695.
11. Renli, W., D. Yueming, and D. Liming. *The application of Apriori-BSO algorithms in medical records data mining*. in *Information Technology, Networking, Electronic and Automation Control Conference, IEEE*. 2016. IEEE.
12. Cheng, J.L. and L. Fang, *Intelligent Data Mining and Knowledge Discovery [M]*. Xi'an Electronic Science and Technology University Press, 2006: p. 240-247.
13. Li, X., Y. Zhu, and X.Z. Luo, *Research and improvement for apriori algorithm of association rule mining.* Computer Knowledge and Technology, 2009. **5**(19): p. 5084-5085.
14. Beltrán-Alzate, C., et al., *Leprosy Drug Resistance Surveillance in Colombia: The Experience of a Sentinel Country.* PLoS Negl Trop Dis, 2016. **10**(10): p. e0005041.
15. Costa, L.G., D. Cortela, and E. IGNOTTI, *Factors associated with the worsening of the disability grade during leprosy treatment in Brazil.* Lepr Rev, 2015. **86**: p. 265-272.

Mobile Health Nutrition Book Design to Prevent Stunting at Childreen <5 years

Vilda Ana Veria Setyawati
Public Health Department
Dian Nuswantoro University
Semarang. Indonesia
vilda.setyawati@dsn.dinus.ac.id

Bambang Agus Herlambang
Informatic Engineering
Universitas PGRI
Semarang. Indonesia
bambangherlambang@upgris.ac.id

Through the national health work meeting, stunting is one of three health problems that are prioritized to be handled holistically by several ministries. However, the treatment are preventive and curative from one party only (the government). Laymen, often do not understand things related to stunting itself. Stunting is a growth disorder that affects children's growth and development. Efforts can be made to support the program by providing health education for mothers. In the digital era, almost everyone has a smartphone that can be used for health education media. This study aims to compile a mobile book design that is expected to help mothers to monitor children's growth and development. The approach used in this study is an object-oriented approach. The method used is the method of developing SDLC prototyping models with usability testing. The steps taken are Requirement Analysis, Design & Development, and Application and Testing. Based on the results obtained, the mobile health nutrition book can be produced and used on mobile phones that have an Android form plate.

Keywords : Health nutrition, mobile book, SDLC, prototyping

I. INTRODUCTION

Toddlers are children aged 0-59 months. Infancy is a golden period in the process of growth and development. In its time Growth and development will be decisive in the success of child growth and development in the next period. Growth is a process of increasing the size / dimensions of the body due to the increasing number and size of cells. Development is the process of organ function maturation, which is shown by the development of abilities, intelligence and behavior [1].

Factors that influence the growth and development of children after birth are poverty, infectious diseases, reduction or deprivation of rights, injustice, limited access to health service resources, food and nutrition information, and ignorance of the relationship between food and health or body needs [2].

Based on Riskesdas 2013 there was an increase in stunting children from 36.8% in 2010 to 37.2% in 2013. Over the past 20 years, handling stunting problems has been very slow. Globally, the percentage of children who are stunted has declined by only 0.6% per year since 1990. WHO proposes a global target of a decrease in the incidence of stunting in children under five years of age by 40% by 2025, but it is predicted that only 15-36 countries meet the target [3].

Health problems that inhibit the growth and development of children, namely stunting. Even according to the results of the national health work meeting (RAKERKESNAS), stunting is one of the priorities besides tuberculosis and improving the quality of immunizations. Cross-sector handling, involves many components and costs are not small. However, being a problem in the region, many people do not understand stunting [4]. One example when stunting was the topic of discussion in a regional head debate, there were several regional heads who were still not exposed to this term. Therefore, education is needed for the community to understand stunting.

Children have the right to foster care, compassion, and sharpening with a maximum age of gold. These three things will have a good impact on the growth and development of children if given appropriately according to their age and sex. People who play a role in child development are mothers. Generally mothers provide education to their children since they were born. Mother is the first and most important teacher for children. A housewife is not a nutritionist, but must be able to compile and assess the dishes that will be served to family members. Knowledge of health nutrition for mothers is needed so that children get proper care. Lack of knowledge and misconceptions about food needs and food values are common problems. One of the causes of stunting problems is lack of knowledge about nutrition or the ability to apply that information in daily life [5].

Previous studies have tested the effectiveness of nutritional education in several ways. Like books, counseling, counseling. But in these ways, the media used for nutritional education is no longer used after the research is completed. Expectations from researchers, media for nutritional education is also used onward for companions in monitoring children's growth and development. So it is necessary to use nutritional education media in accordance with the character of the mother today. The special feature is that it cannot be separated from the gadget.

Mother's character is increasingly following the times. At present, almost all mothers already have cellphones that can be used as health education media. The purpose of this study is to compile a mobile health book that can be used as

reference material for mothers to provide care, compassion, proper teaching so that children avoid stunting.

II. HEALTH EDUCATION

Nutrition and health education is an educative approach to produce improvement or to maintain good health status. The objectives of nutrition and health education are as follows:

a. Can form a positive attitude towards health.
b. Knowledge and skills in selecting and using food ingredients.
c. The formation of good eating habits
d. There is motivation to find out more about things related to health.

Nutrition education in particular to increase the knowledge of mothers aims to change wrongdoing which results in the danger of malnutrition. Media for nutrition and health education is equally important in the process of delivering health information. This media serves as an extension tool. Based on its function, the media is divided into 3, namely [6]:

a. Print media, consisting of: booklets, leaflets, flip charts, rubrics, posters
b. Electronic media is a medium for delivering health information through television, radio, video, or slides, smartphones
c. Media boards (bill boards) usually installed in public places can be used as a medium to convey health messages / information.

III. METHOD

The approach used in this study is an object-oriented approach. In addition, the SDLC method prototyping model is used with usability testing. The steps taken in the prototyping model include planning by performing Requirement Analysis, Design & Develoment Application and Testing (Testing) to obtain returns from prospective users. These stages can be done overlap and cycle.

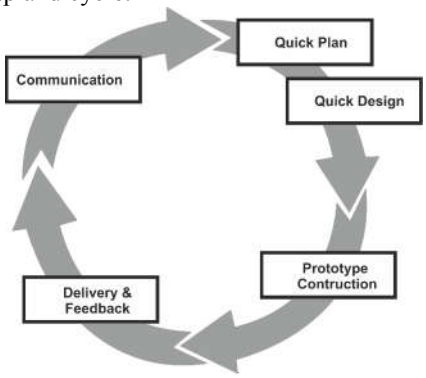


Figure 1: Prototyping Model Diagram

At the planning stage, the requirements analysis is done by using interviews and observations, the program is conducted for prospective application users about what functions are desired from this mobile book application system, in the category of Functional analysis, namely the needs related to the functions obtained by the application system prospective users are expected to be able to display the material / title selection menu, search for titles and there is a menu showing content content. From the results of functional analysis, the researcher conducts a Development requirement analysis which includes the need for tools in building a mobile book application system. Tools used in building application systems include Android Studio 3.0 for development and UML for modeling. At the design stage, the design of a general system design is done with Use Case Diagrams, Activity Diagrams and Class Divisions. In this stage also created a User Interface Design to facilitate application users in knowing the menu layout and layout of this mobile book application system. Some displays that are designed and built include Splash Screen, home page, and pages that display material content. The prototype construction was built using Android Studio 3.0. Usability testing is done to obtain feedback from potential users. As for some things that are tested, the user is asked to test the menu button function, usabiliy testing scenario is done by the user is asked to try to select the material title to see whether the content of the title can appear on the screen or not. Prototype is carried out by usability test for users with criteria for Men / Women aged 25-35 years and familiar with smartphone usage. The results of the feedback of potential users in conducting usability testing are that most prospective users want the menu display that was originally designed with text display in a row list to be displayed with the tile model with an image icon to facilitate the selection of titles and display of material content with images to facilitate understanding of the material. From the results of the feedback, a user interface design is created and builds prototype construction with the tile menu model and the addition of images to material content as desired by the prospective user.

IV. RESULT

The system design tool used is UML (Unifield Modeling Language) which is a graphical language for documentation, specifications, visualization, and building software systems. UML provides standards in notations and diagrams that can be used to model a system [7].

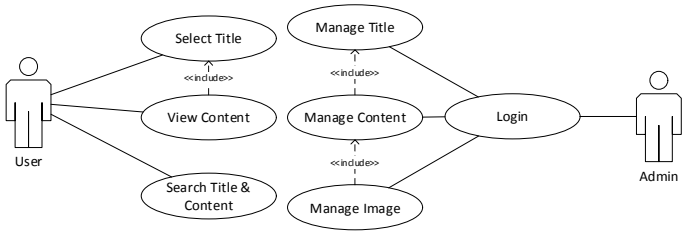


Figure 2 : Use Case Diagram

Use Case Diagram

The use case diagram above serves to explain the interactions that occur in the application system, the series of interactions between users and application systems that occur for more details will be described in the use case scenario as follows.

Nama Use Case : Select Title	
Function	Select the selection of titles provided in the menu and sub menu according to the theme in the mobile book prevention stunting
Use Case Scenario	Users on the main menu page first select a title or subtitles from the menu icon. To be able to see the content the user must choose a title in which there are several options such as Toddler Growth and Development, Stunting, Sharpening, Care and Fostering
Actors Involved	User

Nama Use Case : View Content	
Funtion	Display the contents of the material that has been selected by the user
Use Case Scenario	After the user chooses the title, the user will be able to directly read the content of each title that has been selected.
Actors Involved	User

Nama Use Case : Search	
Fungsi	Look for Titles in mobile book
Use Case Scenario	Users / users can search the content based on the desired title
Actors Involved	User

Activity Diagram

Activity diagram is a depiction of the work flow of the E-Book stunting prevention application. Activities described in the activity diagram are not activities carried out by the user but activities carried out by the system from what the user does. The activity diagram is as follows:

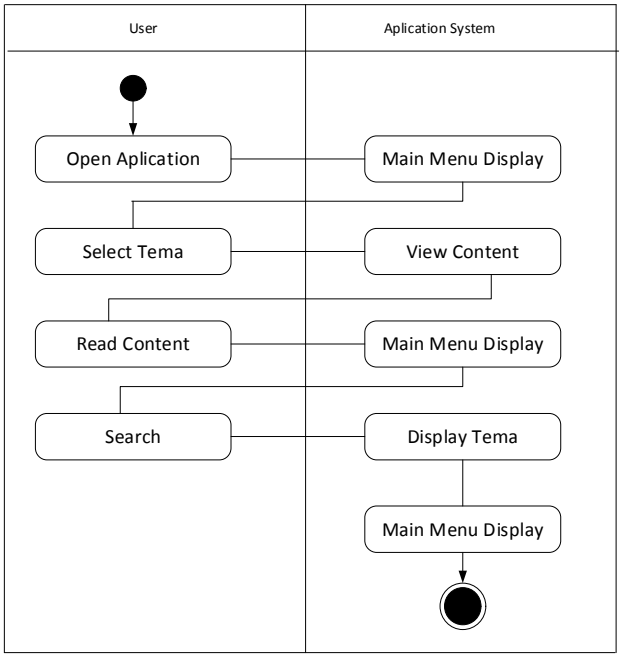


Figure 3 : Activity Diagram

In the activity diagram, the user first opens the application then the system application will display the Main Menu E-Book Prevention Stunting, User then selects the theme or title to be read, from the user system application options will then display the content of the theme or title that the user has selected . The user reads the content that has been selected, the application system provides facilities for searching the title or theme that the user wants and will display search results performed by the user.

Class Diagram

The relationships between classes that occur in the design of the stunting prevention e-book application system can be described as follows:

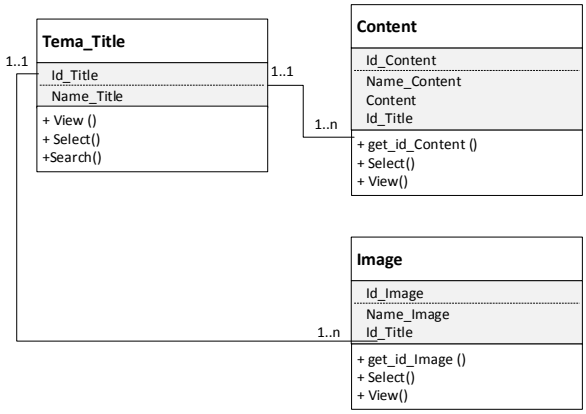


Figure 4 : Class Diagram

Menu Design

Menu design is used to facilitate searching when running a stunting prevention book mobile application program will be created, the menu structure design on the mobile-based mobile book prevention stunting application program can be described as shown above. The menu consists of the main

Menu (Theme / Material, Sharpening, Asih, Foster and Search). In the material menu there are sub menus, among others, toddler growth, development and development, stunting (prevention and impact of stunting) and the right 3A.

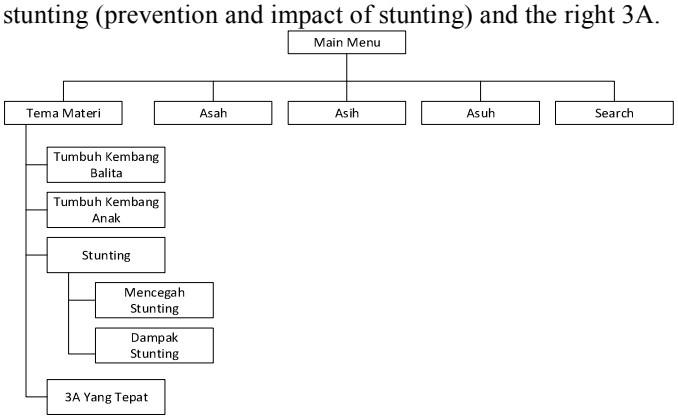


Figure 5 : Menu Design

Implementation Result

Is the result of the steps that have been taken before. The results of the making of the stunting prevention e-book can be seen in the picture below as a sample user interface, which includes the display of the main menu display material and the appearance of the stunting material.



Figure 6: User Interface View

Black Box Testing

Black box testing is conducted to determine the functional validity of the mobile book application system that was built, the test results can be seen in the following table:

No	Testing Scenarion	Expected result	Testing conclusions
1	Users click the material menu button on the main menu	List material with the menu tile model	Valid
2	Users click the sub menu of the Toddler, Grow & Grow material	Show material for User, Toddler & Grow	Valid
3	Users click the stunting material sub menu	Show stunting material content complete with slider model image	Valid
4	User clicks the 3A	Show 3A material	Valid
5	User clicks the Asih menu on the main menu	Appearance icon icon	Valid
6	User clicks on the menu icon	Look at the content of the original material complete with the slider model image	Valid
7	User click the menu on the main menu	Show the sharpening menu icon	Valid
8	Users click the menu icon	Appear the sharpening material content complete with the slider model image	Valid
9	Users click the Foster menu on the main menu	Show Foster menu icon	Valid
10	User clicks on the foster icon	Show foster material content complete with slider model image	Valid

V. DISCUSSION

There are different ways through which parental education can affect their children's health status. Education might have a direct impact on child health because it increases the ability to acquire and process information. This helps parents to prevent stunting and make better health investments for themselves and their children and may result in better parenting in general. Alternatively, education can affect child growth and development through indirect pathways. An increased level of education can give access to more skilled work with higher earnings and these resources could be used to invest in growth and development to cushion the impact of adverse health shocks [8].

Health education is needed by parents as early as possible. If necessary health workers should provide health education since the mother is still pregnant. According to research conducted in Nigeria, maternal education provided since the fetus is still in the womb, effectively increases knowledge and changes health attitudes ($p < 0.05$) [9].

The development of technology in the era of globalization is very rapid as it is today, technology is used to help human work, where in this case mobile-based applications are used to facilitate or help users to find out nutritional status. Based on these conveniences, mobile-based applications are currently very popular because they are easy to use and easy to access anytime and anywhere.

The development of information and communication technology (ICT) today has penetrated into various aspects of

life. The more affordable the prices of hardware such as computers, gadgets and other devices, supported by the increasingly easy operation of software that works on various hardware platforms so that it can help solve various complex computations or use various applications for science, education, economics, entertainment and etc.

In another category of smartphone functionalities, Hardyman et al investigated the use and perceived value of a smartphone with a library of medical textbooks preloaded onto a micro secure digital card to support workplace learning [10].

VI. CONCLUSION

Based on the development method of SDLC prototyping model, the design has been successful through all three processes, namely needs analysis, design and development, and testing. Based on the results obtained, the mobile health nutrition book can be produced and used on mobile phones that have an Android platform.

VII. REFERENCES

[1] A. K.E and L. Marotz, Profil Perkembangan Anak : Prakelahiran Hingga Usia 12 Tahun, Jakarta: INDEKS, 2010.

[2] D. Pudjianto and Kristiani, "Kemiskinan, kondisi geografis dan falitias pelayanan kesehatan yang berhubungan dengan kejadian kekurangan energi proten (KEP) pada balita di Kabupaten Sragen," Universitas Gajah Mada, Yogyakarta, 2013.

[3] K. K. R. Indonesia, "Rapat Kerja Kesehatan Nasional," Kementerian Kesehatan Republik Indonesia, Jakarta, 2018.

[4] B. P. d. Pengembangan, "Riset Kesehatan Dasar "RISKESDAS"," Kementerian Kesehatan Republik Indonesia, Jakarta, 2013.

[5] A. Djaeni, Ilmu Gizi untuk mahasiswa dan profesi Jilid I, Jakarta: Dian Rakyat, 2000.

[6] N. Soekidjo, Pendidikan dan Perilaku Kesehatan, Jakarta: Rhineka Cipta, 2010.

[7] J. Hermawan, Membangun decission support, Yogyakarta: Andi, 2005.

[8] Case, Lubotsky and Paxson, "Economic status and health in childhood : The origin of the gradient," *American economic review,* pp. 1308-1334, 2002.

[9] C. Nwankwo, P. Ezenduka, E. Chiejina and S. Igwilo, "Effect of maternal health education on the health attitude of pregnant motherss attending antenatal clinics in Anambraa State, Nigeria," *IJIRSET,* vol. 5, pp. 1914-1923, 2016.

[10] W. Hardyman, A. Bullock, A. Brown, S. Ingram and Stacey, "Mobile technology supporting trainee doctors's workplace learning and patient care : an evaluation," *BMC Med Education,* vol. 13, 2013.

Improvement of Fuzzy C-Mean using Local Laplacian Filter for Image Segmentation

Gansar Timur Pamungkas
Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
gansar.pamungkas@gmail.com

Rifqi Majid
Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
rifqi.majid29@gmail.com

Guruh Fajar Shidik
Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
guruh.fajar@research.dinus.ac.id

M. Arief Soeleman
Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
arief22208@gmail.com

Ricardus Anggi Pramunendar
Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
ricardus.anggi@research.dinus.ac.id

Ahmad Zainul Fanani
Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
a.zainul.fanani@research.dinus.ac.id

***Abstract*— the FCM (Fuzzy C-Mean) is on of algorithms used in the background image separation research. This study aims to improve the quality of image segmentation using FCM algorithm with Local Laplacian Filter. Thus, we applied the Local Laplacian Filter into the V (Value) and S (Saturation) component of the image. The result of Laplacian Filter then clustered using FCM method. The result of segmentation using FCM and Local Laplacian Filter will be compared with ground-truths images to reveal the value of the PSNR (Peak Signal to Noise Ratio) and the MSE (Mean Square Error). This study also compared the results of the MSE and PSNR with the FCM and K-Mean algorithms. The image is done preprocessing first by using Local Laplacian Filter which gives the color contrast to the image. So the image becomes sharper and when the image changed to the color space HSV already looks quite striking color differences between objects with the background. The results of this study indicate that segmentation using FCM and Local Laplacian Filter has the best MSE and PSNR results compared to the 2 tested algorithms.**

Keywords—FCM, HSV, Image Segmentation, Local Laplacian Filter

I. INTRODUCTION

Image segmentation is one of the methods in digital image processing used to classify an image. Image segmentation goal is to separate an image into many separated regions with every region corresponding to a visual essential object [1]. Image segmentation is used as a fundamental role in many computer vision applications [2].

Implementation of image segmentation can be utilized for various fields of science such as health, economics, medicine, etc. Image segmentation aims to cluster images in accordance with some similar features, i.e. texture, color, intensity, movement, etc. To implement the segmentation method on image, foreground color should be initialized first in RGB and converted the background color to 0 (zero) value for which the distance label is less than 128 [1]. Image segmentation can also applied in video proccessing by comparing each frames to get a movement object that captured [2]. In its development there are many methods used to implement image segmentation [3], one of them is using Fuzzy C-Mean algorithm [4].

The FCM (Fuzzy C-Mean) is one of methods implemented in the field of image classification and has some advantages, which is widely used in various research and goals. Based on the statement of H. Pei, et al [5], the FCM is an algorithm implemented in the soft clustering, which is based on fuzzy set theory. This algorithm can classify the image by changing the object of the image into the HSV color space [6]. HSV color space is better than the RBG to representing human view, because the HSV has a color range from 0 to 1 [7]. In this color space the FCM algorithm clusters the imagery based on the similarity of color and texture.

In this research the role of preprocessing the image becomes very important to get better image segmentation quality. Therefore the authors propose a method to improve the results of image segmentation using FCM so as to get maximum results. Local Laplacian Filter (LLF) is a method of color image correction that makes the color of the image becomes sharper [8]. The Laplacian of an image is a secondary method of enhancement [9].

Refer to Guangyu Zhang [8] using Local Laplacian Filter as an approach in which quickly make a high quality line drawings. In his research, he construct a local intensity remapping and then construct it into Laplacian Pyramid by computing a Laplacian pyramid of the transformed image.

Satish Bhairannawar [9], in his research, before implementing the Laplacian method he convert the color space from RGB to HSV. He applying the Laplacian filter to Value and Saturation component. To obtain the enhanced Value by stretching the contrast and appropriating the Saturation component, the output of luminance component is subjected to CLAHE (Contrast Limited Adaptive Histogram Equalization). And the last was reconverting from HSV back to RGB.

In this research, authors thought that LLF method is very suitable to be combined with FCM algorithm to get better image segmentation result. As mentioned above that FCM classify HSV component and LLF is stretching V and S component. To test the results of image segmentation using the authors' proposed model, the authors used sample dataset images obtained from Berkeley Segmentation

Dataset (BSD) [10].

Referring from the information above, research that using FCM combined with LLF as a filter is not yet investigated. Applying Laplacian filter to V and S component. Then cluster it using FCM using the image datasets to produce a better result of image segmentation obtained by using the results of the calculation of PSNR (Peak Signal to Noise Ratio) and MSE (Mean Square Error). Those are often implemented to evaluate the image segmentation results along with its ground truth. From their process, the value then calculated, the lower the value of MSE can be indicated that it is a better image segmentation. On the other hand, the higher the PSNR value, indicates a better image segmentation.

Hence, for the whole process, the MSE and PSNR results through the proposed model will be compared with the results of the FCM and K-Mean algorithms that are widely used in research to perform image segmentation.

II. FUNDAMENTALS

A. Image Segmentation

Image segmentation is widely applied in various researches that have many challenges in the field of computer vision [2]. Those researches most often implementing some fast and effective algorithms, which can easily be applied in the image segmentation field.

For several reasons, the main objective of the image segmentation research is to group some pieces of image into regions through its similar properties (i.e. intensity, color, spatial statistics) purposed to create the meaningful image and is easier for the analysis process [1]. The grouping of blocks that associated with similar image features can be called as the result of segmentation. Then, the most significance steps to achieve the successful segmentation is through grouping some schemes and choosing the image features.

B. Fuzzy C-Mean Clustering

It is said that the FCM is a well-recognized technique and categorized as a Hard K-Means [11]. The FCM is implementing some model of the fuzzy grouping to set up the level or degree of membership in a range at the interval of 0 (zero) to 1 (one) [4]. To count the membership level in a class or a cluster is determined by its membership degree. Here is the formula given (1):

$$Tfc = (max(xi\ (y{=}1) + min(xi\ (y{=}2))) / 2 \quad (1)$$

Determinating the center of the cluster become the basic concept of FCM method. Restoring repeatedly every membership values and the cluster center will show the center of the cluster precising its location [12]. The effectiveness and easy implementation is the reason for using the general algorithm of image segmentation using FCM. The function membership piece of fuzzy by the pixel of an image to the convinced amount of classes. To start the observed intensity collection in a pixel j, we used the equation (2):

$$yi = [yj1,yj2,....yjN]T \ , j = 1,.....,M \quad (2)$$

where yi, (i = 1,...., N) (in the *RMI*, N is the sum amount of *TR/TE*), and M is represent the pixel position sum number. The spectral canal amount can be the pixel intensity of the spectral canal inserted to the N. The minimum iteration between the membership function of fuzzy *uj* and the center of pixel vector in every single class *vk* is calculated in equation (3):

$$| \text{FCM} = \sum_{ij} \sum_{k} u_k (i,j) - V_k \quad || \quad (3)$$

where $Uk(I,j)$ is the membership amount in the pixel location (i,j) in the k class is. The q is the content exponent existed in every membership of single fuzzy determined as parameter and the fuzziness amount of the classification outcome. To reflecting the degree of similarity between the class center point and the pixel vector of the mentioned location, the value of 8 memberships is ranged from 0-1. Measuring how close the value function to 1 is providing the minimum FCM. The further the pixel intensity from the center is the smaller membership value it get [3].

C. Local Laplacian Filter

The LLF (Local Laplacian Filter) is one of techniques pf digital image processing, which is implemented often for color image enhancement. The LLF counts the result by calculating the LP (Laplacian Pyramid) independently each layer [9]. Once the Laplacian pyramid computed, the output image can easily be obtained by expanding the LP [9].

The LP is composed with spatially constant Gaussian kernels. It is said that the LP is widely recognized as the poor method to the edge-aware operations and some edges positioning [13].

Particularly, the Laplacian Filter is efficient to find the fine detail in an image. Some researchers conducted previously to train the Laplacian operator to find the enhancement of any feature along with the sharp discontinuity. In this study, to approximate the second derivative in the Laplacian operator, we used equation (4).

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \quad (4)$$

Before stepping to further processing, the Laplacian filter is implemented to count the S (Saturation) and V (Value) components, which is then used to sharpen the contents of the image. The CLAH (Contrast Limited Adaptive

Histogram) is implemented to process the enhanced luminance component. By applying the Laplacian filter to the V and S components will give the output, shown as V_α and S_α [9].

In this research, Local Laplacian filter used for increase the contrast from the dataset images [10]. The result of image segmentation processing was affected by quality of preprocessed images. After that, preprocessed images would be converted from RGB to HSV, then image will be clustering processed using Fuzzy C-Mean algorithm.

III. Research Experiment

This study aims to segment the image to distinguish the object image with background. The following sequence of research process conducted (1):

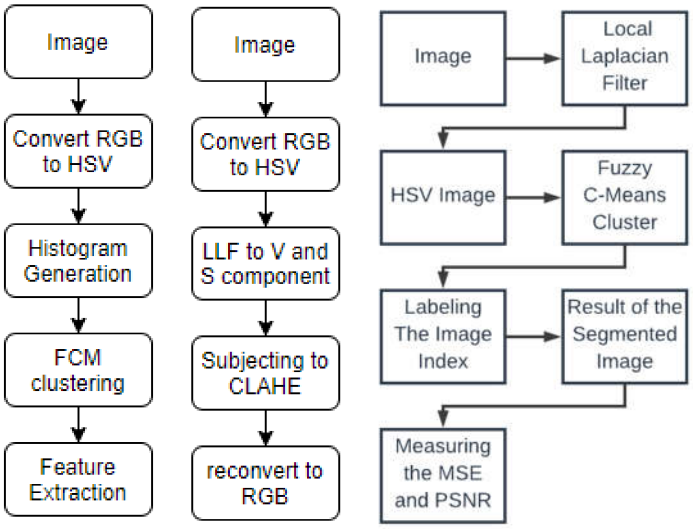


Fig. 1. (a) Research Flow Diagram of Implementing FCM on Image Segmentation by Jie Zhang, (b) Flow Diagram of Implementing Local Laplacian Filter by Satish Bhairannawar, (c) Flow Diagram of Proposed model by Adding LLF before clustering using FCM

Fig 1. (a) shows how the authors of that research implementing FCM on image segmentation, Fig 1. (b) shows hoe the authors of that research implementing Local Laplacian Filter to sharper value of V and S components. In this research, the authors implementing Local Laplacian Filter before FCM to get a better image segmentation result.

The original image to be processed will be enhanced by using Local Laplacian Filter, which aims to make image contrast better. For that implementation, the image will be changed firstly from RGB (Red, Green, Blue) into the HSV (Hue, Saturation, Value) color. That changing is clustered using the Fuzzy C-Mean technique by determining the center of the cluster and entering the Hue and Saturation values of the image. The Figure (2) shows the segmentation result.

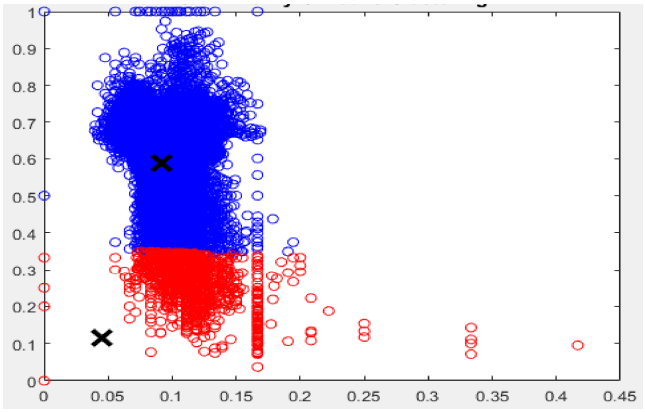


Fig. 2 Graph shown clustering using Fuzzy C-Mean

In the graph (2) it looks Fuzzy C-Mean algorithm successfully segmented to degrade object and background image. Image sequence result of image segmentation process can be seen in figure (3) following:

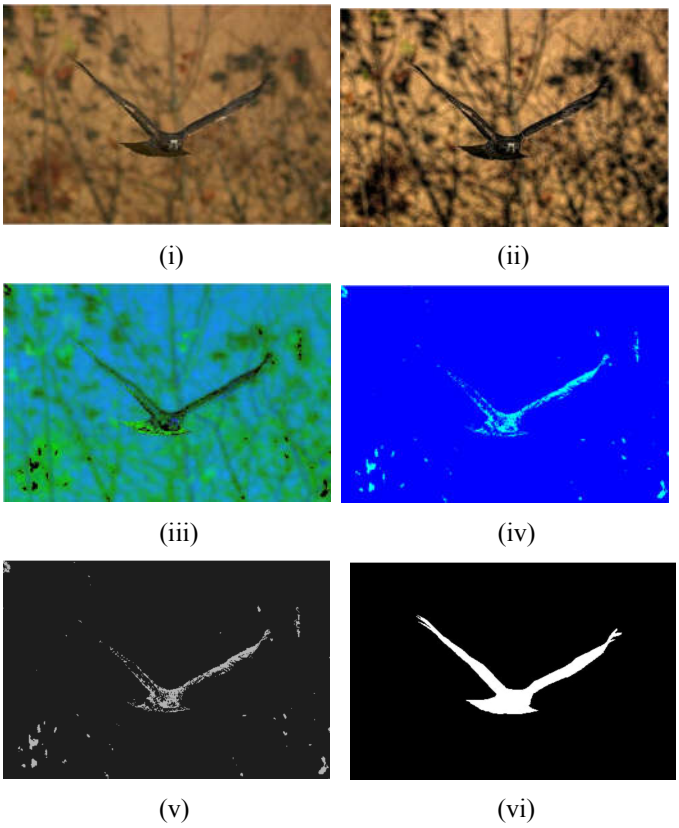


Fig 3. The (i) is the original image, (ii) is the image after processing the Local Laplacian Filter, (iii) is the image in HSV, (iv) is the result of segmented image, (v) is the segmented Image in grey, (vi) is the Ground-Truth Image

In figure (3) shows the process in this research from the selecting dataset images, preprocessing image using local laplacian filter, and then the result is image clustering using Fuzzy C-Mean.

After the image successfully segmented the next step is to compare the image quality of the results of segmentation with ground-truth image. In this study, we use the PSNR (Peak Signal to Noise Ratio) and the MSE (Mean Square Error) to measure the proposed model quality. The MSE

calculation formula can be seen at equation (5) as well as the PSNR calculation is at (6).

$$MSE = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}[f(i,j) - f_1(i,j)]^2 \quad (5)$$

$$PSNR = 10\log_{10}\left(\frac{255^2}{MSE}\right) \quad (6)$$

MSE and PSNR were selected for this research because previous studies [14] were used to monitor and evaluate the results of image segmentation. Furthermore, this research will compare the value of MSE and PSNR of two algorithms done in previous research with our proposed model. The author compares the results of MSE and PSNR from 10 Dataset Image with 2 other algorithms previously used, Fuzzy C-Mean and K-Mean. Here's the comparison:

TABLE I. COMPARISON OF MSE AND PSNR RESULT USING FCM, K-MEAN, AND PROPOSED MODEL

Image	Fuzzy C-Mean		K-Mean		Proposed Model	
	MSE	PSNR	MSE	PSNR	MSE	PSNR
1	28116,7711	3,6411	36001,7632	2,5676	25344,2251	4,0920
2	41755,8675	1,9236	40581,1299	2,0476	41253,5554	1,9762
3	37545,8165	2,3852	39799,6490	2,1320	26310,4235	3,9295
4	36682,2648	2,4862	27126,4599	3,7969	28556,1138	3,5738
5	39740,7150	2,1384	33225,0531	2,9161	33298,1311	2,9066
6	32238,1705	3,0471	26461,7863	3,9046	22336,7905	4,6406
7	31466,1085	3,1524	32828,3458	2,9683	29976,4122	3,3630
8	38566,0196	2,2688	33521,2318	2,8776	25828,7365	4,0098
9	38202,6286	2,3099	28748,8528	3,5446	27483,8490	3,7400
10	39315,7581	2,1851	27916,6052	3,6722	37717,7195	2,3653
Average	**36363,0120**	**2,5538**	**32621,0877**	**3,0428**	**29810,5957**	**3,4597**

Based on the experimental result shown in table (1), the authors compare 10 (ten) sample images obtained from the public datasets [15] and compare the results of the segmentation using 3 (three) algorithms, i.e. Fuzzy C-Mean, K-Mean, and FCM + Local Laplacian. The MSE results can be seen in the following figure (5).

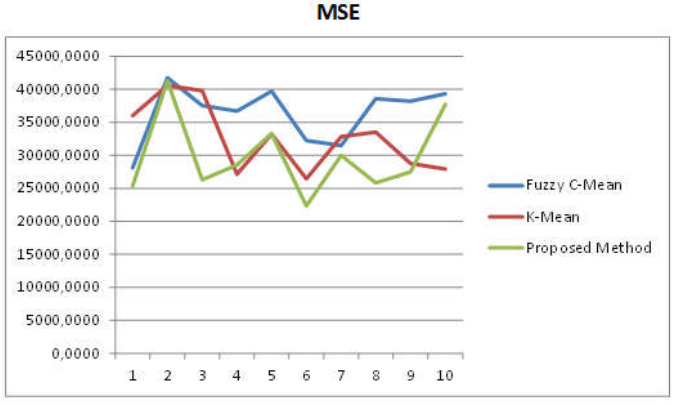


Fig 5. Graph of MSE Results from 3 Algorithms

From the result of MSE in figure (5) it is seen that the proposed model (Fuzzy C-Means + Local Laplacian Filter) has a lower error than the 2 algorithms tested with an average value of 29810.5957 compared with the average value of FCM and K-Mean that is equal to 36363,0120 and 32621,0877. The results show that the implementation of the fuzzy C-Means along with the incorporation with Local Laplacian Filter create the best MSE value compared to other 2 (two) algorithms. Furthermore, this study will make the graph of PSNR value comparison with 3 methods that were tested before. For graph of PSNR result can be seen in figure (6) :

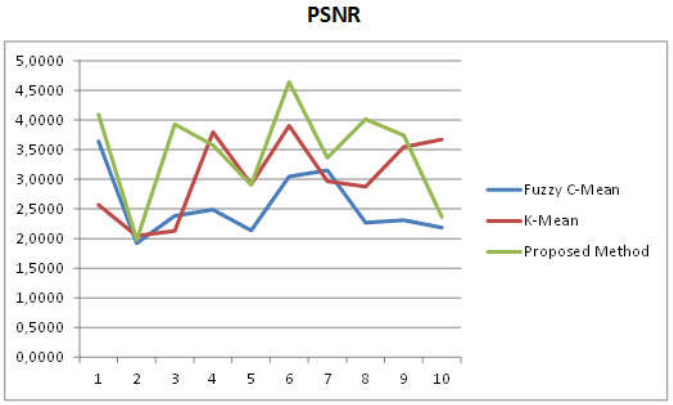


Fig 6. Graph of PSNR Results from 3 Algorithms

From the result of PSNR in figure (6) shows that the proposed model has the highest PSNR value with an average of 3.4597 compared with the mean value of FCM and Kmean that is 2,5538 and 3,0428. From these results it can be concluded that the image of the segmentation result using FCM and Local Lapaclian Filter has the best PSNR value compared to the 2 tested algorithms. Comparison of the segmented image between proposed model with FCM and K-Means can be seen in figure (7) below :

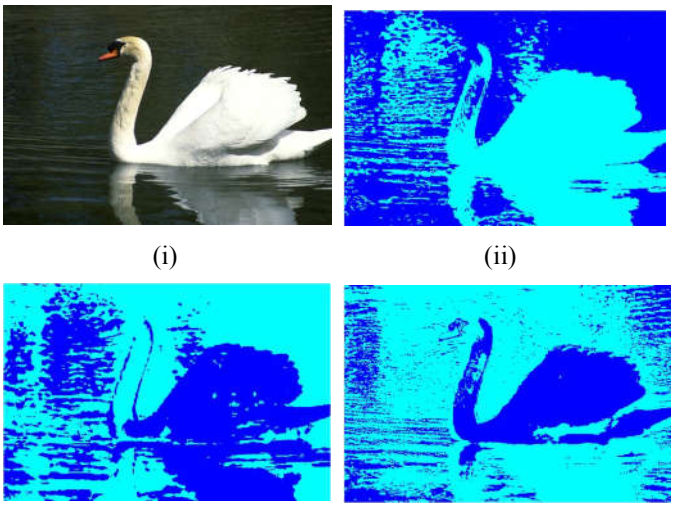


Fig 7. (i) is the original image, (ii) is the segmented image using FCM, (iii) is the segmented image using K-Means, (iv) is the segmented image using the proposed model.

IV. CONCLUTION

This research has been successful to show improvement of segmentation result using Fuzzy C-Mean and Local Laplacian Filter algorithm compared with 2 algorithms tested, that is using Fuzzy C-Mean and K-Mean. The results of this study provide the best value of MSE and PSNR compared to the 2 algorithms tested because before the image is clustered using Fuzzy C-Mean algorithm, the image is done preprocessing first by using Local Laplacian Filter which gives the color contrast to the image. So the image becomes sharper and when the image changed to the color space HSV already looks quite striking color differences between objects with the background. This results is shown the better method of image segmentation rather than the Fuzzy C-Mean and K-Mean algorithms. With this research the authors hope the future research can be developed into a method that can increase the quality of detection or classification of an object image and bring much benefits for various fields of science.

REFERENCES

[1] R. A. Pramunendar, C. Supriyanto, Dwi Hermawan Novianto, Ignatius Ngesti Yuwono, G. F. Shidik, and P. N. Andono, "A classification method of coconut wood quality based on Gray Level Co-occurrence matrices," in *2013 International Conference on Robotics, Biomimetics, Intelligent Computational Systems*, 2013, pp. 254–257.

[2] F. N. A. Guruh Fajar Shidik, "Time Frame Selection Based Feature Extraction for Fire Detection in Video Surveillance," *J. Comput. Theor. Nanosci.*, vol. 20, no. October 2014, pp. 1885–1889, 2014.

[3] W. Wang and C. Wu, "Image segmentation by correlation adaptive weighted regression," *Neurocomputing*, vol. 267, pp. 426–435, 2017.

[4] S. K. Choy, S. Y. Lam, K. W. Yu, W. Y. Lee, and K. T. Leung, "Fuzzy model-based clustering and its application in image segmentation," *Pattern Recognit.*, vol. 68, pp. 141–157, Aug. 2017.

[5] H. Pei, Z. Zheng, C. Wang, C.-N. Li, and Y. Shao, "D-FCM: Density based fuzzy c-means clustering algorithm with application in medical image segmentation," *Procedia Comput. Sci.*, vol. 122, pp. 407–414, 2017.

[6] J. Zhang and M. Du, "Fast Image Segmentation of Gold Immunochroma atographic Strip Based on o FCM Clustering Al lgorithm in HSV Color Space," no. Cisp, pp. 525–528, 2012.

[7] G. F. Shidik, F. N. Adnan, C. Supriyanto, R. A. Pramunendar, and P. N. Andono, "Multi color feature, background subtraction and time frame selection for fire detection," *Proc. 2013 Int. Conf. Robot. Biomimetics, Intell. Comput. Syst. ROBIONETICS 2013*, no. November, pp. 115–120, 2013.

[8] G. Zhang, J. Lin, X. Gao, and M. Liao, "A Fast Local Laplacian Filter for High Quality Suggestive Contours Generation," no. Cisp, pp. 313–318, 2015.

[9] S. Bhairannawar and A. Patil, "Color Image Enhancement using Laplacian filter and Contrast Limited Adaptive Histogram Equalization," pp. 0–5, 2017.

[10] J. Z. Hui Li, Jianfei Cai , Thi Nhat Anh Nguyen, "A Benchmark For Semantic Image Segmentation." [Online]. Available: http://www.ntu.edu.sg/home/asjfcai/Benchmark_Website/benchmark_index.html.

[11] A. H. Rangkuti, Z. E. Rasjid, and D. J. Santoso, "Batik Image Classification Using Treeval and Treefit as Decision Tree Function in Optimizing Content Based Batik Image Retrieval," *Procedia Comput. Sci.*, vol. 59, pp. 577–583, Jan. 2015.

[12] K. Mahesh and K. Kuppusamy, "A New Hybrid Video Segmentation Algorithm using Fuzzy C Means Clustering," *J. Comput. Sci.*, vol. 9, no. 2, pp. 229–237, 2012.

[13] S. Paris, S. W. Hasinoff, and J. Kautz, "Local Laplacian filters: edge-aware image processing with a Laplacian pyramid," *Acm Siggraph 2011*, vol. 58, no. 3, pp. 81–91, 2015.

[14] M. A. Soeleman, M. Hariadi, and M. H. Purnomo, "Adaptive threshold for background subtraction in moving object detection using Fuzzy C-Means clustering," *TENCON 2012 IEEE Reg. 10 Conf.*, pp. 1–5, 2012.

[15] C. Fowlkes, "Figure-Ground Assignments in Natural Images," 2017. [Online]. Available: https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/fg/.

DATASETS

To test the results of image segmentation using the authors' proposed model, the authors used sample dataset images obtained from Berkeley Segmentation Dataset (BSD) [15]. Below is the dataset used for this research and its result from proposed model.


Image Dataset 1

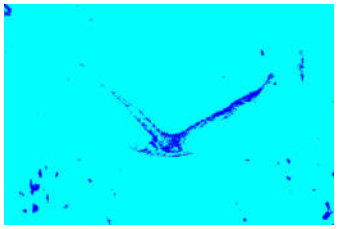

Segmentation Result 1


Image Dataset 2

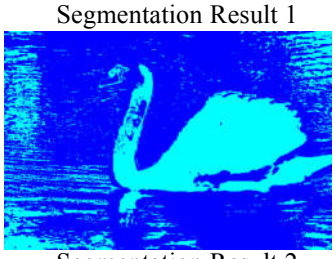

Segmentation Result 2


Image Dataset 3

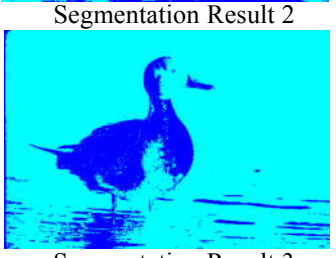

Segmentation Result 3


Image Dataset 4

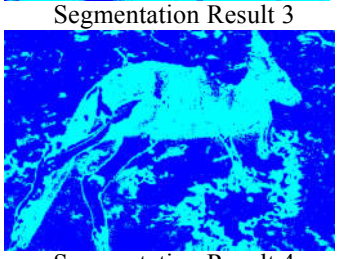

Segmentation Result 4


Image Dataset 5

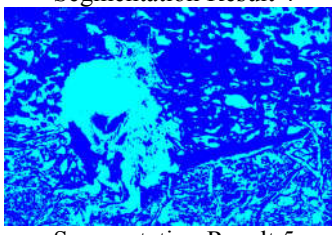

Segmentation Result 5


Image Dataset 6

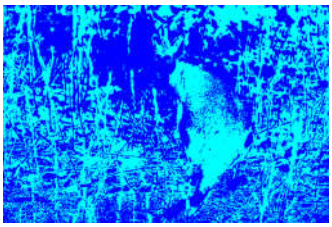

Segmentation Result 6


Image Dataset 7

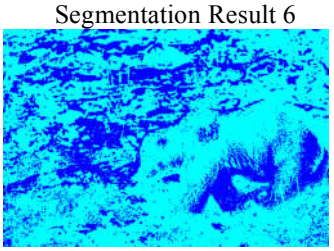

Segmentation Result 7


Image Dataset 8

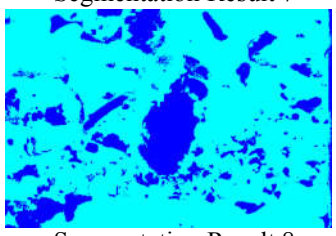

Segmentation Result 8


Image Dataset 9

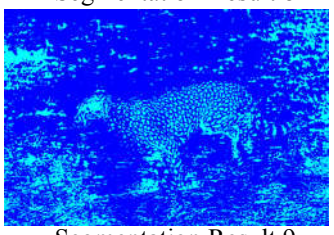
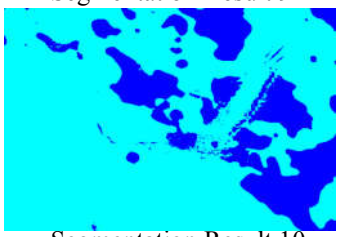

Segmentation Result 9


Image Dataset 10


Segmentation Result 10

Performance Comparative of AODV, AOMDV and DSDV Routing Protocols in MANET Using NS2

Alamsyah[1,2], Eko Setijadi[1]
[1]Dept. of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[2]Dept. of Electrical Engineering, Universitas Tadulako Palu, Indonesia
E-mail: alamsyah14@mhs.ee.its.ac.id, ekoset@its.ac.id

I Ketut Eddy Purnama[3,4], Mauridhi Hery Purnomo[3,4]
[3]Dept. of Electrical Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
[4]Dept. of Computer Engineering, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
E-mail: {ketut, hery}@its.ac.id

***Abstract*—mobile ad hoc network (MANET) is a network that has fast characteristics, efficient deployment, and can apply in an emergency location such as natural disasters, military operations, forest fires, and health monitoring. However, the challenges and problems facing MANET are dynamic topology changes, with no fixed infrastructure, and limited energy sources. Dynamic topology changes will result in redundant data packets on each node, and the quality of the network will decrease. To overcome problems in MANET then selection of routing protocol is essential in improving network quality. This study aims to compare the performance of AODV, AOMDV, and DSDV routing protocols based quality of service (QoS). Parameters QoS to be analyzed are packet delivery ratio (PDR), packet loss, throughput, delay, and routing overhead (RO). The simulation result using NS-2 shows that AOMDV performs better than AODV and DSDV regarding PDR. AODV performs better than AOMD and DSDV regarding throughput, RO. While DSDV performs has better than AODV and AOMDV regarding the delay.**

Keywords—MANET; AODV; AOMDV; DSDV; QoS

I. INTRODUCTION

The use of wireless networks in recent years has become popular and attractive to investigate, especially in MANET. MANET utilization has widely used in areas with no fixed infrastructure, such as communications in military operations, flood detection, earthquake detection, forest fire detection, and health monitoring. Examples of health monitoring applications such as the use of heartbeat sensors using the ZigBee protocol. Other than that, MANET has a characteristic that is capable of managing its own built systems and the rapid spread of communication [1]. However, some of the problems that MANET has in communicating, such as dynamic topology changes, no fixed routers [2], and limited energy sources due to battery usage [3].

The problem of topology which dynamic will result in overloading of packets in each neighboring node [4]. To overcome the changing topological and occurrence of excessive packet delivery in each adjacent node, the selection of routing protocols in MANET is crucial in improving QoS such as PDR, throughput, packet loss, delay and routing overhead.

Some researchers have proposed the MANET field using different parameters and simulations. The performance evaluation of the DSDV and AODV in MANET using NS2 based on changes in number and node speeds indicates that AODV performs better than DSDV regarding PDR, delay, and throughput [5]. DSDV performance is superior to AODV regarding energy consumption [6].

The comparison between the DSDV and AOMDV routing protocols using the NS2 is proposed [7]. In this paper, the authors explain the difference between spectrum sensing performance and route discovery on AOMDV and DSDV based on delay parameters, PDR.

Analysis of the AODV and DSR protocols based on the number of nodes and speed indicate that AODV superior to DSR. While DSR excellent to AODV regarding less dense nodes [8]. The choice of random waypoint and random walk mobility are recommended using the DSR protocol in the improve PDR and throughput [9].

Analysis of the AODV, DSR, and DSDV on MANET indicate that DSR is superior to AODV regarding PDR, energy efficiency, and delay [10].

The literature study of the AODV, DSR, and AOMDV using the NS2 shows that AOMDV superior to AODV and DSR in finding alternative routes. However, AOMDV incurs more routing overhead when broadcasting to each neighboring node. DSR performs excellently than AODV and AOMDV regarding PDR, throughput [11].

Analysis of the AODV with considering the number of nodes indicates that AODV has better PDR performance at ten nodes and better RO performance at 50 nodes. [12].

The performance of AODV, DSDV, and OLSR for health monitoring shows that AODV supports to applied to small-scale networks. OLSR supports to apply to large-scale networks [13].

Analysis of AODV, AOMDV, DSR, and DSDV with consideration on nodes change and pause time. The simulation results using NS2 indicates that AODV is better than AOMDV, DSR, and DSDV regarding CBR connections. DSR is superior to AODV, AOMDV, and DSDV regarding TCP connections.

However, routing overhead (RO), throughput and UDP connections have not been evaluated [14].

This study analysis of AODV, AOMDV, and DSDV with consideration of the QoS. The parameters used in analyzing QoS are PDR, throughput, packet loss, delay, and RO. The discussion of this research described as follows. Section 1 MANET explanation in general and research related to MANET topics. Section 2 routing protocol in MANET. Section 3 simulation and analysis. Section 4 conclusion of the research results obtained.

II. ROUTING PROTOCOLS ON MANET

MANET is an ad-hoc network technology that can adapt to mobile conditions. MANET grouped into three routing protocols, namely reactive, proactive, and hybrid [15]. Fig. 1 shows the routing protocol on MANET.

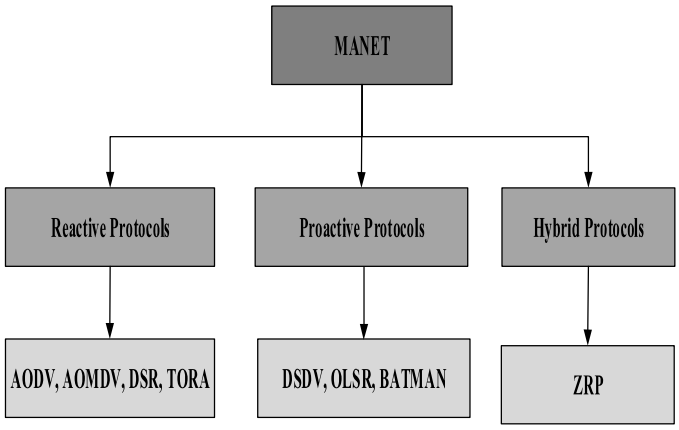


Fig. 1. Routing protocols on MANET

A. AODV

This protocol is the reactive routing protocols that only requests a route when needed and maintain the timer-based state at each node according to the use of the routing table [16]. A node considers the path as an active path, if it sends, accepts or passes packets to that route. So each intermediate network node must make decisions regarding packet data transmission. The route discovery packet widely distributed to the source who wants to reach the destination. Furthermore, the network topology changes should be sent only to the node that will need the information. The problem of AODV is not to support asymmetric relationships. That is, AODV is capable of supporting only symmetric connections between nodes. AODV creates a route using route request (RREQ) and route reply (RREP).

B. AOMDV

This protocol is a multipath-based AODV protocol development and has many characteristic equations with the AODV protocol [17]. AOMDV is vector based and hop by hop. Besides, AOMDV provides two primary services: route discovery and route maintenance. The main difference between AODV and AOMDV is the number of routes found in each route search or route discovery. However, AOMDV can correct fault routes and determine multiple paths to the destination. The weakness of AOMDV is during the process of route discovery of many overhead messages that occur. This overhead message occurs due to the response of some RREQs.

C. DSDV

DSDV is protocols development from bellman-ford routing. The DSDV routing concept, every network topology change automatically the routing table is updated regularly [18]. The changes of network topology make the node detect and send the route update packets to its neighboring node. The routing table that each node contains: the destination address, the number of hops needed to reach the destination, and the sequence number. Before sending updates, nodes wait completion time to ensure it does not receive updates from its old neighbors.

III. RESULT OF SIMULATION AND ANALYSIS

The simulations we use in analyzing the performance of AODV, AOMDV, and DSDV NS-2. The analysis quality of service (QoS) on routing protocols using the AWK script and written in OTCL language. The simulation area used is 1000 m x 1000 m

The simulation scenarios used are the area of simulation, packet size, simulation time, number of nodes, simulation speed. The propagation and mobility model uses two-ray and random motion. The performance analysis of AODV, AOMDV, and DSDV uses five parameters consisting of PDR, packet loss, throughput, delay, and routing overhead (RO).

TABLE I. SIMULATION SETUP

Parameters Name	Description
Simulator	NS 2.35
Operating System	Linux Ubuntu 14.04
Routing Protocols	AODV, AOMDV, DSDV
Number of Node	25, 50, 75, 100, 150, 200
Radio Propagation	Two-Ray Model
Packet Size	512 bytes
Transmission Type	802.11
Speed	20 m/second
Mobility Model	Random Motion
Simulation Time	200 seconds
Simulation Area	1000 m x 1000 m

A. PDR

PDR is one of the parameters in QoS to indicate the success rate of a routing protocol. Fig. 2 shows that the movement of PDR values on AODV and AOMDV tends to increase. However, the change of PDR values in DSDV tends to decrease. The increase of PDR on the AOMDV protocol is due to the ability of this protocol in determining the route and fixing some route error from the source to the destination. AOMDV performance is superior to AODV and DSDV regarding PDR. The mean value of PDR in AOMDV is 60.09%, AODV is 57.42%, and DSDV is 38.86%.

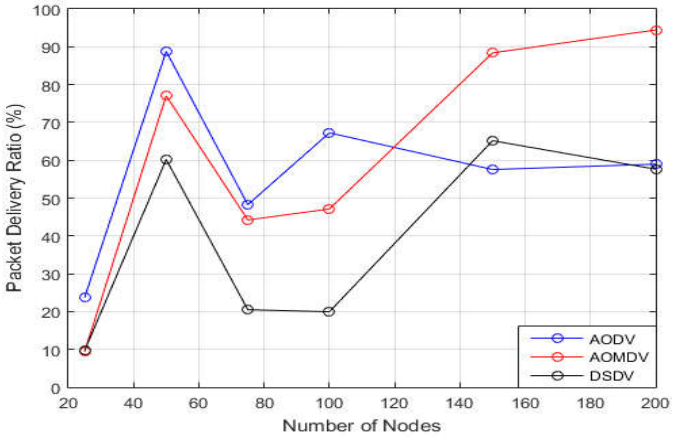


Fig. 2. PDR vs number of nodes

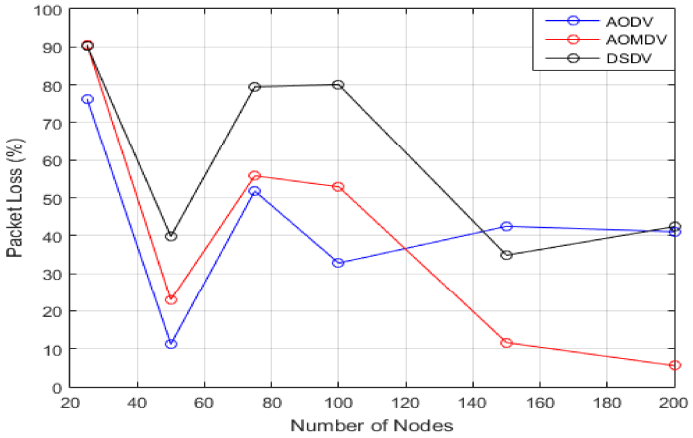


Fig. 4. Packet loss vs number of nodes

B. *Throughput*

Throughput describes the condition of the data rate in a network. Fig. 3 shows the throughput performance based on the number of node densities. The movement of throughput values in AODV, AOMDV, and DSDV is unstable and tends to decrease. However, the average throughput performance on AODV is better than AOMDV and DSDV. Average AODV throughput of 368 Kbps, AOMDV of 278.5 Kbps, and DSDV of 201.42 Kbps. The average increase in throughput of AODV is due to AODV-owned characteristics that only request a route when needed.

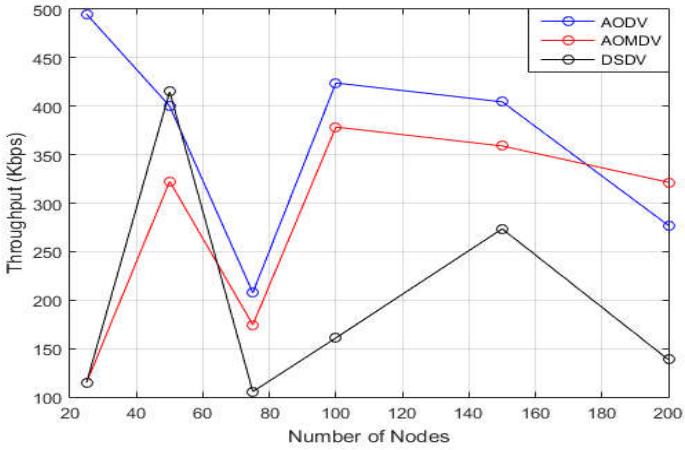


Fig. 3. Throughput vs number of nodes

C. *Packet Loss*

This parameter occurs due to a queue that exceeds the buffer capacity of each node. Fig. 4 shows the movement of AOMDV packet loss on the number of more dense nodes that tend to fall. For AODV and DSDV, the change of the resulting packet loss values tends to enhance in denser nodes. A decrease in packet loss on AOMDV indicates that the number of packets lost to the destination is very small compared with AODV and DSDV. The average AOMDV packet loss is 39.9%, AODV is 42.58%, and DSDV is 61.14%.

D. *Delay*

Delay is sending packets from source to destination within a time interval. Fig. 5 shows the delayed movement on AODV, AOMDV, and DSR tends to decrease in each node addition. However, DSDV performance superior to AODV and AOMDV regarding delay parameters. The decrease in the delay value generated by the DSDV protocol is due to the characteristics of the routing table and the delivery of packets to all neighboring nodes working periodically or structured. Performance average delay on DSDV is 27.8 milliseconds, AOMDV of 61.19 milliseconds, and AODV of 102.83 milliseconds.

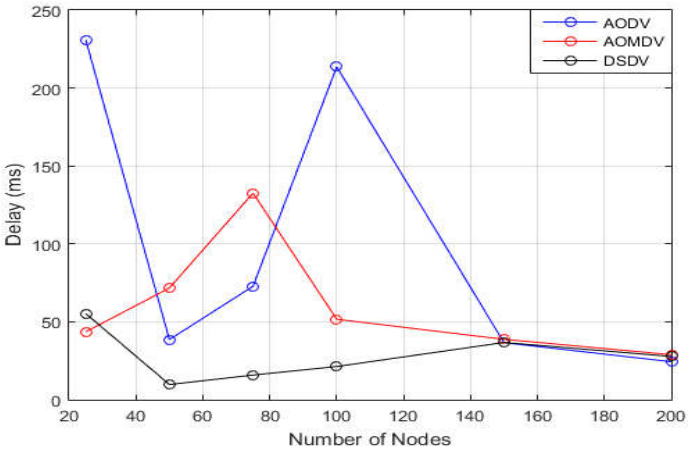


Fig. 5. Delay vs number of nodes

E. *Routing Overhead (RO)*

The RO parameter shows the working efficiency of a routing protocol. A high RO value will result in bandwidth wastage in packet delivery. Figure 6 shows the movement of RO on AOMDV and DSDV tends to rise. For the change of RO values on AODV tends to decrease in each addition of nodes. So it can be said that RO performance on AODV is better than DSDV and AOMDV. Average RO performance on

AODV of 1809 packet, DSDV of 139586 packets, and AOMDV of 171464 packets.

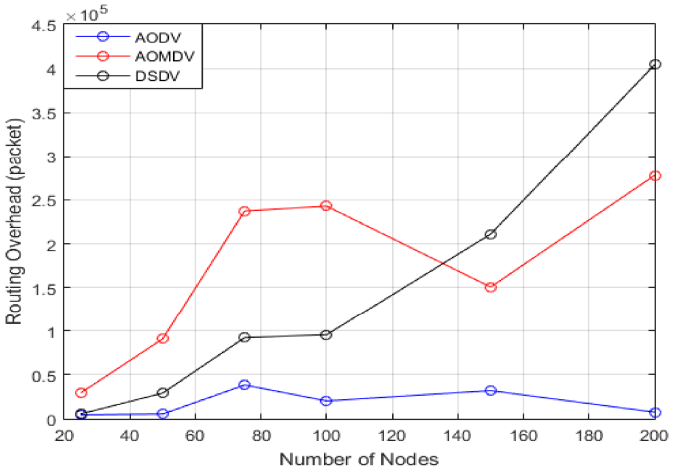


Fig. 6. RO vs number of nodes

IV. CONCLUSION

This study aims to analyze AODV, AOMDV, and DSDV based on QoS by considering node density. The simulation results show that AOMDV performance superior to AODV and DSDV regarding PDR, packet loss. AODV performs better than AOMDV and DSDV regarding throughput, RO. DSDV has better than AODV and AOMDV regarding the delay. In this study, it can say that for smaller numbers of nodes it is recommended to use the AODV protocol. The more significant number of nodes is supported using the AOMDV protocol.

REFERENCES

[1] B. Xu, S. Hischke, and B. Walke, "The role of ad hoc networking in future wireless communications," in Proc. of the IEEE International Conference on Communication Technology (ICCT), vol. 2, pp. 1353-1358, 2003.

[2] W. Abdou, A. Henriet, C. Bloch, D. Dhoutaut, D. Charlet, and F. Spies, "Using an evolutionary algorithm to optimize the broadcasting methods in mobile adhoc networks," Journal of Network and Computer Applications, vol. 34, no. 6, November 2011, pp. 1794–1804.

[3] V. Rishiwal, A. Kush, and S. Verma, "Stable and energy efficient routing for mobile adhoc networks," IEEE Fifth International Conference on Information Technology: New Generations (ITNG 2008), April 2008, pp. 1028-1033, pp. 1028-1033.

[4] F. Tekiner and A. Shrestha, "On MANET routing protocols for mobility and scalability." In International Conference on Parallel and Distributed Computing, Applications and Technologies, p.p. 451-456, IEEE Computer Society, November 2009.

[5] A. Daas, K. Mofleh, E. Jabr, and S. Hamad, "Comparison between AODV and DSDV routing protocols in mobile ad-hoc network (MANET)," IEEE National Symposium on Information Technology: Towards New Smart World (NSITNSW '15'), February 2015.

[6] L. M Doddamani, S. M Yaragop, A. Chikaraddi, and S. Kanakaraddi, " Energy consumption comparison of AODV and DSDV routing protocols," IEEE International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT '16'), March 2016, pp. 2128-2132.

[7] A. Mehta and A. Gupta, "Retrospection and comparison of DSDV and AOMDV routing protocols in Manet Using Ns-2," IEEE International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT'14'), February 2014, pp. 325-329.

[8] O. Kembuan, Widyawan, and S. S. Kusumawardani," Analisis kinerja reactive routing protocol dalam mobile ad-hoc network (MANET) menggunakan NS-2 (Network Simulator)," JNTETI, Vol. 1, No. 1, Mei 2012.

[9] P. Nayak and P. Sinha,"Analysis of random waypoint and random walk mobility model for reactive routing protocols for MANET using NetSim simulator," Conference AIMS 15, 2015, pp. 427-432

[10] S. E Khediri, N. Nasri, A. Benfradj, A. Kachouri, and A. Wei, "Routing protocols in MANET: Performance comparison of AODV, DSR and DSDV protocols using NS2," IEEE International Symposium on Networks, Computers and Communications (SNCC'14'), June 2014.

[11] M. Y. Barange dan A. K. Sapkal, "Review paper on implementation of a multipath reactive routing protocol in MANET," Conference ICEEOT *16*. 2016, pp. 227-231.

[12] B. Karthikeyan, N. Kanimozhi, and S. H. Ganesh,"Analysis of reactive AODV routing protocol for MANET," Conference WCCCT 14, 2014, pp. 264-267.

[13] Alamsyah, M. H. Purnomo, I. E Purnama, and E. Setijadi, "Performance of the routing protocols AODV, DSDV and OLSR in health monitoring using NS3," Conference IEEE ISITIA 16, 2016. pp. 323-328.

[14] B. Paul, K. A. Bhuiyan, K. Fatema, and P.Pratim Das. , "Analysis of AOMDV, AODV, DSR, and DSDV routing protocols for wireless sensor network*,"* IEEE International Conference on Computational Intelligence and Communication Networks (ICCICN), Nov 2014, pp. 364-369.

[15] Ashwani Kush and Sunil Taneja, "A Survey of routing protocols in mobile ad hoc networks," International Journal of Innovation, Management, and Technology, Volume 1, No. 3, pp. 279-285, August 2010.

[16] A. Bagwari, R. Jee, P. Joshi, and S. Bisht, "Performance of AODV routing protocol with increasing the MANET nodes and its effects on QoS of mobile ad hoc networks," Conference IEEE CSNT 12, 2012, pp. 320-324.

[17] M. K. Marina and S. R. Das, "Ad hoc on-demand multipath distance vector routing," Wireless Communications and Mobile Computing, pp. 969-998, 2006.

[18] M. Kumar, R. and Mishra, 'An Overview of MANET: History, Challenges and Applications", Indian Journal of Computer Science and Engineering (IJCSE), vol. 3, no. 1, pp. 121-125, 2012.

Bit Localization in Least Significant Bit using Fuzzy C-Means

Atika Mutiarachim[1], Stefano Felix Pranata[2], Basirudin Ansor[3], Guruh Fajar Shidik[4], Ahmad Zainul Fanani[5], Arief Soeleman[6], Ricardus Anggi Pramunendar[7]

Master of Informatics Engineering, Faculty of Computer Science, Dian Nuswantoro University
Imam Bonjol street 205-207, Lt 2, Semarang. Telp. (024) 3547038, (024) 70793727
E-mail : amutiarachim@gmail.com[1], stefanofelixpranata@gmail.com[2], anzhorhmm@gmail.com[3], guruh.fajar@research.dinus.ac.id[4], ahmad.zainul.fanani@dsn.dinus.ac.id[5], arief2802@dsn.dinus.ac.id[6], ricardus.anggi@dsn.dinus.ac.id[7]

Abstract—Least Significant Bit (LSB) as one of steganography methods that already exist today is really mainstream because easy to use, but has weakness that is too easy to decode the hidden message. It is because in LSB the message embedded evenly to all pixels of an image. This paper introduce a method of steganography that combine LSB with clustering method that is Fuzzy C-Means (FCM). It is abbreviated with LSB_FCM, then compare the stegano result with LSB method. Each image will divided into two cluster, then the biggest cluster capacity will be choosen, finally save the cluster coordinate key as place for embedded message. The key as a reference when decode the message. Each image has their own cluster capacity key. LSB_FCM has disadvantage that is limited place to embedded message, but it also has advantages compare with LSB that is LSB_FCM have more difficulty level when decrypted the message than LSB method, because in LSB_FCM the messages embedded randomly in the best cluster pixel of an image, so to decrypted people must have the cluster coordinate key of the image. Evaluation result show that the MSE and PSNR value of LSB_FCM some similiar with the pure LSB, it means that LSB_FCM can give imperceptible image as good as the pure LSB, but have better security from the embedding place.

Keywords— Clustering, Fuzzy C Means, Image Processing, Least Significant Bit, Steganography

I. INTRODUCTION

Tremendous development of technology is directly proportional with the more risk of technology crime that may arise. That's why security aspect also be concern research search trend in this year [1]. Secure factor is needed to dodge a risk, even more in this IoT era, security system be a big deal and big challenge to solve [2]. There are many crime case that may happened inside technology like theft of important and secret message. To handle it, many researchers vying to find the best of encryption method, based on the time reduce, memory capacity improvement, and the hassle of method in embedding and decrypting the messages. Messages (especially a secret messages) tend to be privacy and usually contains of some important information so should not be receive by unauthorized parties [3].

Information hiding is one of security systems method to embed a message into an object like video, audio, image, fingerprint and etc. This method consist of steganography and watermarking. Steganography divided into linguistic and technical. In watermarking there are robust and fragile. Each method has own characters, advantages and disadvantages. The choice of method should be based on the need of security systems that will be created.

In this research, LSB was combined with Fuzzy C-Means (FCM) to optimize the difficulty level to decrypt the message from an stego-image. FCM is a clustering method, that can cluster one data to two or more cluster, based on their characteristic [4]. In this research, FCM used to cluster the pixel of an image. By using FCM, message will embedded to a random part of an image. This methods aim to looking for a cluster with big number of pixel as place of message, and make the embedded message more secure. It is a novel method for security system, especially in steganography. In future, CM_LSB also can give an improvement as contribution to Internaet of Thing (IoT) security systems.

II. RELATED WORKS

To increase insight, researcher start with analyze and study some related works trend in steganography. Some papers that significantly contributed in this research are summarized in this section.

First the idea of reversible data hiding, [5] use difference expansion and difference histogram shifting method to get high capacity and reversibility in embedding data. Another research is about hybrid concept of data security using LSB and Discrete Cosine Transform (DCT), aims to optimize the security and authentication of data. DCT is an algorithm that used to compress signal and image data [6]. Cryptography and QR code are used to combine the LSB and DCT [7].

In [8] pure LSB compared with LSB-Canny edge detection method to ensure the embedding message process runs perfectly and increase the imperceptible between the original image and the embedded one. Using four *.bmp grayscale image, [9] *.txt message embedded to the image, then be invaded by JPEG compression, gaussian noise, salt and pepper and media filtering, then measured by PSNR and

MSE. The measuring result show that LSB-Canny method get better value of PSNR and MSE than pure LSB.

The next method, [10] use LSB Matching Revisited method (LSBMR) with Sobel as the edge detection of an image. By choosing two pixels from the cover image randomly using pseudo random number generator (PRNG), then embedded the bits on those pixels using stego-key. Not all pixel suitable to adjusted, because the sharp change may be visible to human eye [11]. LSBMR applied to overcome those limitation, by use the edge to embedded information first, then leave the smooth areas based on the capacity of messages. Compared with Enhanced Least LSBMR (EALSBMR), the development of LSBMR, based on PSNR evaluation value [10] LSBMR has a slightly better quality in image and in resisting visual attack than EALSBMR, because LSBMR not adjusted the smooth regions.

III. MATERIAL AND METHOD

Three approaches to design a security system are choose applicable image, select appropriate embedding location and choose best encrypted version [12]. Embedding location of messages is get from clustering result of FCM. Encrypted method that use here is LSB. There are five images that used in this research, with different size and dimension. The image order is start from the left that is rainbow cube as first image, and the right that is stroberi as fifth image.

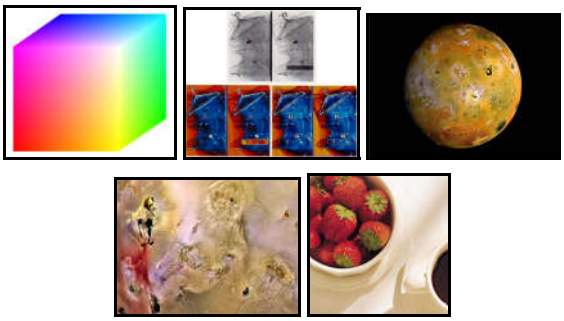


Fig. 1 Five Cover Images

A. *Proposed Method*

There are many security systems research trends in information hiding. Especially in steganography, using basic principles of LSB algorithm, many study expand this algorithm or combined it with many kinds embedded place of detector method like Sobel, Canny and etc. This research focused on technical digital image steganography, and also expanding the LSB method to be LSB_FCM. In this research, a novel parallels method was applied by combine LSB and FCM. Framework of this research is in figure 2.

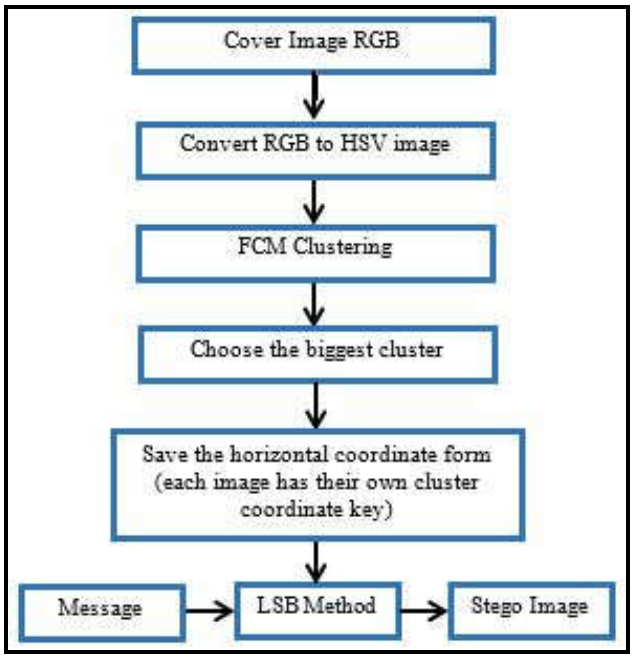


Fig. 2 Framework

FCM will cluster the image, then choose the best cluster with biggest capacity, save the cluster coordinate key to embeded the message using LSB. Unspecified cluster help to make the message more secure, because more difficult to guess than if embedded in each edge of image.

B. *Fuzzy C-Means*

Fuzzy C Means (FCM) is a clustering method which able to grouped a datasets into n cluster, with each data point belonging to each cluster in certain degree. FCM have ability to minimize error in clustering function [5], and the formation of new clusters have close membership values to existing class [13]. Here FCM used to cluster the cover image pixels, then choose the best cluster with more amount of pixel as embedded region.

C. *Least Sigificant Bit*

Least Sigificant Bit (LSB) is a steganography method which are simple in embedded the bit [14] [15], and imperceptible [9], so there is no contrast and different look beetwen the original image and the embedded image . It's used to embedding the message into the picture, by put the bit of a message to the last bit of each picture pixel.

One RGB pixel, consist of R, G and B value. The value of each R, G and B is beetwen 0-255 or 11111111 and 00000000 or combination beetwen 1 and 0 in biner. From eight biner combination, the last bit or eightth bit of each R,G,B in each pixels are used to be the embedded region of messages. The messages that will be embedded to image, must in form of biner too. LSB is easy to implemented, high perpetual transparency and imperceptibilty because the cover image usually look same with the embedded image.

D. LSB_FCM

LSB_FCM is the combination of embedding message method that use FCM to looking for the best cluster, and use LSB as embedded method. It is a novel combination of data security method, especially in steganography. Application of LSB_FCM in this research as follow:

Input : *1) Read RGB cover image*

Process : *2) Convert RGB to HSV image*
3) Cluster HSV image using FCM, divide image pixels into two clusters
4) Choose the cluster that has biggest capacity
5) Save the cluster coordinate as cluster coordinate key to avoid the coordinate change when the image clustered again. It is also as reference when decrypted the message. Each image has different cluster coordinate key.
6) Embed the message into the choosen capacity coordinaet using LSB method

Output : *7) FCM_LSB stego-image*

Two flowcharts in figure three and figure four show the different process in pure LSB and LSB_FCM.

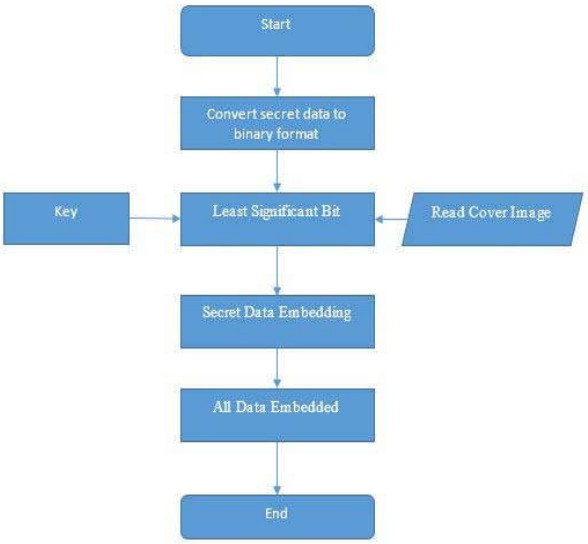


Fig. 3 Flowchart of Pure LSB Process

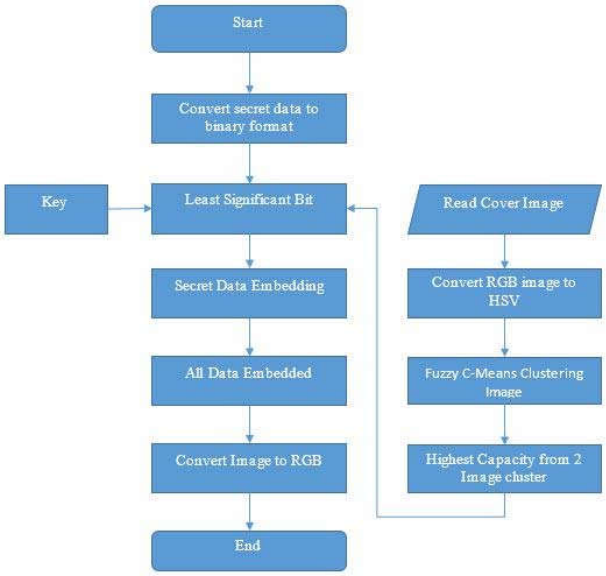
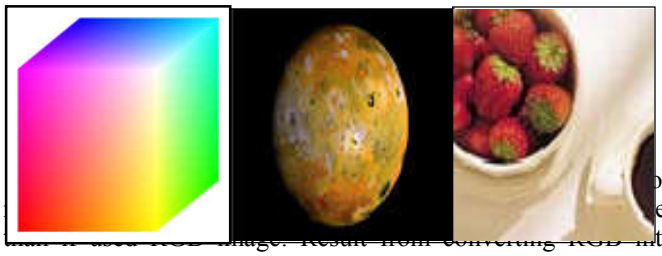


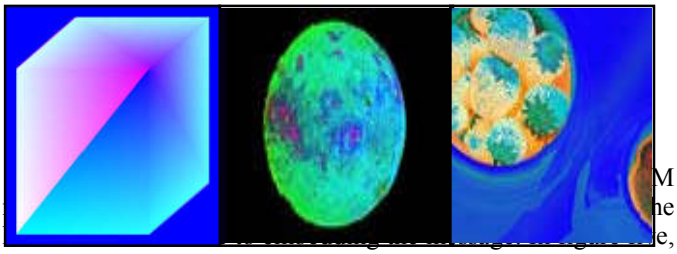
Fig. 4 Flowchart of LSB_FCM Process

IV. RESULT AND DISCUSSIONS

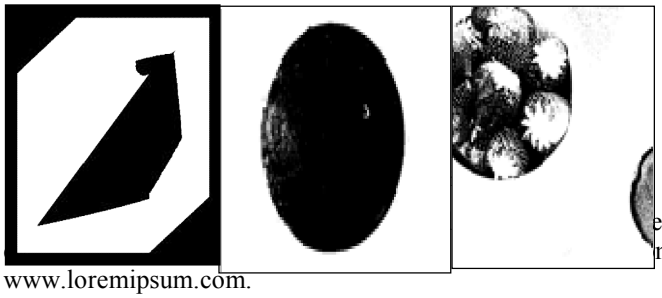
Five standart RGB images are used as the cover images. Below is three from five images that used in this research.



on
er
into
HSV images are displayed in figure four.



M
he
e,
the white part of the image indicate the biggest cluster. The horizontal coordinate from the biggest cluster will save as coordinate cluster key. This key also be the reference to decode the embedded message.



en
m
www.loremipsum.com.

'Lorem ipsum dolor sit amet, consectetur adipiscing elit. Sed dignissim pulvinar purus sit amet rhoncus. Donec iaculis suscipit lectus, nec ullamcorper lacus mollis ut. Sed varius, nulla at fringilla placerat, nunc dolor volutpat purus, nec tempor mauris nulla at sem. Sed consequat mi neque, vitae convallis velit dapibus in. Vivamus blandit nibh ut accumsan viverra. Donec dolor orci, lacinia vel lorem non, vestibulum molestie felis. Proin sit amet nisl iaculis, condimentum nulla vel, convallis massa. Curabitur suscipit nibh nec neque ullamcorper commodo. Mauris nec ex faucibus, tincidunt felis vel, suscipit diam. Mauris interdum sem at lectus porta, hendrerit pretium erat vestibulum. Pellentesque eget hendrerit metus. Nulla sagittis odio velit, sit amet egestas mi mattis in. Mauris sed nunc in elit sagittis mattis pellentesque eu nibh. Maecenas ultricies eros eget tortor mattis accumsan. Cras vulputate sit amet turpis sed blandit. Proin sit amet elit leo. Nulla eu purus nibh. Proin facilisis egestas massa nun.'

Fig. 8 The Message

The stego-images are displayed in figure nine. From the embedded process, the result show that there is no contrast look between the cover images and stego-images, it is the main keys in steganography.



Fig. 9 Stego-images

The bit capacity, byte capacity, iteration, and time execution were calculated from each image. The result show in table I.

TABLE I. LSB_FCM

Image	Bit	Byte	Iteration	Time
1	226089	28261	3	0.32867076
2	142242	17780	10	0.33652582
3	401352	50169	3	0.46154190
4	365709	45713	3	0.37597524
5	1045269	130658	13	0.34054169

There is no big change and significant different in the stego-images when compare by the cover image, it is prove by the MSE and PSNR values. In some images, MSE and PSNR from LSB_FCM give sligthly better quality than the LSB stego-images, especially for the MSE values.

TABLE II. MSE AND PSNR

Image	LSB		LSB_FCM	
	MSE	PSNR	MSE	PSNR
1	0.0097	68.2611	0.0097	68.2443
2	0.0159	66.0902	0.0161	66.0634
3	0.0551	70.7139	0.0055	70.7265
4	0.0058	70.4413	0.0059	70.4116
5	0.0030	73.2787	0.0031	73.2426

V. CONCLUSION

FCM_LSB method has a weakness that is because the data that can be embedded is limited to the bit amount of choosen cluster, it can be smaller than embedded place from original LSB method. This limit also can be an advantages, because if the data just embedded on a particular cluster, so it will be more difficult to decode the message when compared with original LSB. People must know the cluster coordinate key of an image to decode the embedding message. There is no contrast look between the cover image and the steo-image, it is really good in steganography. Compare with LSB, MSE and PSNR of LSB_FCM method s

REFERENCES

[1] K. Panetta, “Gartner Top 10 Strategic Technology Trends for 2018,” *Smarter With Gartner*, 2017. [Online]. Available: https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2018/. [Accessed: 01-Feb-2018].

[2] C. Stergiou, K. E. Psannis, B. G. Kim, and B. Gupta, “Secure integration of IoT and Cloud Computing,” *Futur. Gener. Comput. Syst.*, vol. 78, no. December, pp. 964–975, 2018.

[3] R. M. L. Jr and G. P. Schell, *Sistem Informasi Manajemen*. 2008.

[4] A. Cheddad, J. Condell, K. Curran, and P. Mc Kevitt, “Digital

image steganography: Survey and analysis of current methods," *Signal Processing*. 2010.

[5] J. Nayak, B. Naik, and D. H. S. Behera, "Computational Intelligence in Data Mining - Volume 2," vol. 32, no. November, 2015.

[6] X. Wu, J. Weng, and W. Q. Yan, "Adopting secret sharing for reversible data hiding in encrypted images," *Signal Processing*, vol. 143, pp. 269–281, 2018.

[7] R. Sharma and J. Singh, "Image Authentication Technique Based on Digital Watermarking using Clustering," *Int. J. Adv. Res. Comput. Sci.*, vol. 8, no. 5.

[8] R. K. Singh and D. K. Shaw, "A Hybrid Concept of Cryptography and Dual Watermarking (LSB_DCT) for Data Security." p. 12, 2018.

[9] A. Susanto and C. A. Sari, "PERLINDUNGAN HAK CIPTA PADA CITRA DIGITAL MENGGUNAKAN," vol. 8, no. 2, pp. 441–448, 2017.

[10] Z. Fouroozesh and J. Al Ja'am, "Image steganography based on LSBMR using Sobel edge detection," *2014 3rd Int. Conf. e-Technologies Networks Dev. ICeND 2014*, pp. 141–145, 2014.

[11] R. J. Anderson and F. A. P. Petitcolas, "On the limits of steganography," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 4, pp. 474–481, 1998.

[12] M. S. Subhedar and V. H. Mankar, "Current status and key issues in image steganography: A survey," *Comput. Sci. Rev.*, vol. 13–14, no. C, pp. 95–113, 2014.

[13] M. Asyali, D. Colak, O. Demirkaya, and M. Inan, "Gene Expression Profile Classification: A Review," *Curr. Bioinform.*, vol. 1, no. 1, pp. 55–73, 2006.

[14] M. Pavani, S. Naganjaneyulu, and C. Nagaraju, "A Survey on LSB Based Steganography Methods," *Int. J. Eng. Comput. Sci.*, vol. 2, no. 8, pp. 2464–2468, 2013.

[15] V. Verma, Poonam, and R. Chawla, "An enhanced Least Significant Bit steganography method using midpoint circle approach," *Int. Conf. Commun. Signal Process. ICCSP 2014 - Proc.*, pp. 105–108, 2014.

Pork and Beef Features Extractions

Jaza Chairunnisa[1], Tiaz Iskandar Muda[2], Fari Katul Fikriah[3], Ricardus Anggi Pramunendar[4], Guruh Fajar Shidik[5], Ahmad Zainul Fanani[6], Arief Soeleman[7]

Master of Informatics Engineering, Faculty of Computer Science, Dian Nuswantoro University
Imam Bonjol street 205-207, Lt 2, Semarang. Telp. (024) 3547038, (024) 70793727
E-mail : jazachairunnisa@gmail.com[1], tiaziskandar95@gmail.com[2], farichatulfikriyah45@gmail.com[3], ricardus.anggi@dsn.dinus.ac.id[4], guruh.fajar@research.dinus.ac.id[5], a.zainul.fanani@dsn.dinus.ac.id[6],arief22208@gmail.com[7]

Abstract*—Increased demand for meat in the market causes various problems, including misuse of sales by mixing beef with pork. The similarity between meat as making a layman is difficult to distinguish between meat. Ignorance makes consumers can not distinguish the type of meat that makes consumers just the origin of buying meat. Therefore, the introduction of meat is important to ensure that meat is certain. Color and texture is one of the characteristics of every meat. In this study we used the introduction of meat using images using histogram-based texture extraction and HSI color feature extraction. The result of feature extraction obtained then processed by using Neural Network, SVM, and KNN method to be grouped into two classes of meat, namely pork and beef. The determination of the success of this classification process is evaluated by using a confusion matrix which yields an accuracy of 93.75% using k-fold 20 with Neural Network method for the classification.

Keywords-component; Image Processing, Meat Classification, Neural Network, Feature Extractions

I. INTRODUCTION

The increasing demand for meat in the market to be evidence that meat is one of the foods that are often consumed. The high demand for meat affects the stability of the price of meat itself. In general, the meat is sold merchants according to type. However, not a few traders who do mix to get bigger profit [1]. Differences in the price of beef and pork, often used by traders to do mixing, when sending meat consumer orders. Fraud committed by traders like this, harm consumers to get meat that is not in accordance with the desired [8].

The color of the meat becomes one of the most important quality characteristics for buyers. Red meat is the name for meat that comes from mammals or reptiles that have several types, such as beef, mutton, pork, and other meats [2]. The similarity between meat and meat often makes traders mix meat consumption [3]. This is what makes a layman difficult to distinguish between meat.

There is a characteristic difference between beef and pork, this difference is the underlying study. One way to recognize beef and pork in the field of informatics is to use image processing.

Lestari Handayani, et al. In their study showed the difference between beef and pork based on color and texture. The Hue Saturation Value (HSV) model is used for color extraction and uses the Probabilistic Neural Network (PNN) method for feature classification and extraction. The study also compares three methods: Gabor, Principle Component Analysis (PCA) and Local Binary Pattern (LBP). The experiment was conducted using 100 images of beef and pork and mixed. The best spread value of the ratio of the three methods is 108 [4].

Sachin D Khirade and A.B Patil in their study mentioned that health monitoring and disease detection in plants is very important for sustainable agriculture.A large number of manual jobs require an excessive process of being difficult to monitor disease in plants.In the study discussed the method, some feature segmentation and extraction were used to detect the disease in plants[5].

In another research[6], explored to produce features that can be used and classification system.Different quantization models can produce different accuracy even when compared to more complex methods.In the study yielded superior results for color on feature extraction by using the MSB quantization method.The results in the study show that feature extraction with simplified quantization of images and dimensional reduction can provide more compact vector results.

In the studies [7] conducted experiments using self tuning MLP Classifier (AutoMLP) and SVM for the quality classification of coconut wood.n the study autoMLP resulted in an accuracy of 78.82%, which is slightly better than 77.06% SVM using the SVM classification of libSVM libraries.Research other by Gaurav Kumar and Pradeep Kumar Bhatia[9]. In the study discussed how important in image processing. The research explains that there are various preprocessing techniques of drawing prior to extracting features that will be useful in classifying and recognizing images. The techniques mentioned in the study are such as binarization, thresholding, resizing, normalization etc. Image processing technique applied before first extraction is done. The results of this study concluded feature extraction can be done and ascertain which feature extraction techniques produce the best features to perform extraction based on the type of image complecity.

Research by Mr. Hrishikesh P, et al. Presenting that characteristic of leaf disease is an important thing that will help in finding the right disease on the leaf itself. In that study disease detection with the proposed methodology using leaf features. Hue's image of HSI provides clear and more useful discrimination for size-sizing[10].

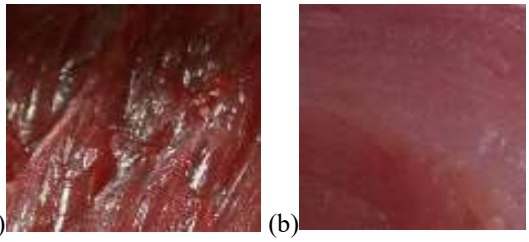
The study used a feature-based histogram texture extraction and color feature extraction of Hue Saturation Intensity (HSI).

The purpose of this research is to get the accuracy of classification of meat image classified in two classes, namely pork meat using Neural Network method with dataset which is the result of histogram extract based on feature texture combined with HSI color extraction result in meat figure

II. RESEARCH METHODS

A. Data Collection Method

Data collection is a stage that aims to obtain information relating to research. In this study, data were used is primary data with 40 images of beef and 40 images of pork taken directly from our laboratory. Images that taken using DSLR cameras are then did preprocessing using Adobe Photoshop to cropping each image has the same size that is 400x400 pixels [16]. For example shown at fig 1.



(a) (b)

Figure 1. (a) beef (b) pork

B. Method

In this research, the image data that has been through preprocessing is RGB image which then extract color features on RGB images into HSI images so that there are 12 color features of each image and feature-based histogram texture extraction that produces 6 texture features of the beef and pork image This feature is used as data classification process Artificial Neural Network. So the value of accuracy obtained from the classification using the evaluation technique of confusion matrix.

C. Proposed Feature Extraction

a. Color Feature Extraction

Image processing has many methods to get the desired features, one of which is color-based extraction features. In this study the color feature extraction used is Hue Saturation Intensity[13].

b. Texture Feature Extraction

In this research, Image Processing used is histogram texture feature method[15], which search for 6 features in the picture, among others mean, standard deviation, entropy, energy, slope, and fluency. The extraction on the image is done by the calculation process as follows on equation 8-13:

1) Intensity Meaning Feature (Mean):

The average intensity or so-called mean is the first feature calculated by the following formula:

$$m= \sum_{i=0}^{L-1} i.p(i) \quad (8)$$

2) Skewness Feature

$$Skewness= \sum_{i=0}^{L-1}(i-m)^3 . \mathrm{p(i)} \quad (9)$$

The skewness feature shows the negative value of the brightness distribution on the left leaning against the mean and the positive value indicates the brightness distribution is tilted to the right against the mean.

3) Standart Deviation Feature

The third feature is to calculate the standard deviation with the following formula:

$$\sigma= \sqrt{\sum_{i=0}^{L-1}(i-m)^2 . \mathrm{p(i)}} \quad (10)$$

4) Energy Feature

$$\text{Energi} = \sum_{i=0}^{L-1}[p(i)]^2 \quad (11)$$

The energy value that is often called the uniformity value in the image has a maximum value of 1. An image that has many gray levels will have little energy value compared to the image that has a small gray level.

5) Entropy Feature

$$\text{Entropi} = \sum_{i=0}^{L-1} p(i)\ \log_2\ (p(i)) \quad (12)$$

Entropy shows the complexity of an image so that the higher the image entropy value the more complex the image will be.

6) Smoothness Feature

$$\text{Smothness (R)} = 1 - \frac{1}{1+Deviasi^2} \quad (13)$$

The highest Smoothness (R) value indicates that the image has good value.

D. Classification

Support Virtual Machine (SVM)

SVM is an algorithm that seeks to find a linear separator (hyperplane) between data points in a multidimensional space of two classes.The benefit of SVM is to create better accuracy than other classifiers. SVM is perfect for handling interactions between features and redundant features. The main problem with SVM is the correct selection of kernel functions. different kernel functions show different results on each data. SVM takes more time than other algorithms and is designed to solve binary class problems. Problems with SVM are not as many as other methods. Easily handle complex nonlinear data points[17].

KNN

The algorithm aims to classify new objects based on attributes and training samples. KNN is very easy to implement and requires relatively fast training time. But the KNN algorithm has its drawbacks as in slow and noise sensitive testing[17].

Neural Network

Neural network (neural network) is a non-linear statistical data modeling tool consisting of simple elements that operate in parallel. Networks are adjusted based on output and target comparisons. Artificial neural networks have the function of pattern recognition, identification, classification and difficult problem solving for conventional or human computers.In addition, neural networks have the advantage of solving complicated cases in spite of time that is not fast[17].

In this study, both features are combined for use in the Neural Network (NN) classification process, so there are 18 features on each image, consisting of 12 features resulting from color extraction, and 6 features extracted histogram features. Feature extraction is obtained by using MATLAB program while the NN classification process uses RapidMiner.

E. Features

X1 = MeanHue
X2 = MeanSaturation
X3 = MeanIntensity
X4 = StDevHue
X5 = StDevSaturation
X6 = StDevIntensity
X7 = SkewnessHue
X8 = SkewnessSatuation
X9 = SkewnessIntensity
X10 = CurHue
X11 = CurSaturation
X12 = CurIntensity
X13 = Mean
X14 = StDev
X15 = Skewness
X16 = Energi
X17 = Entropi
X18 = Smoothness

F. Evaluation

To evaluate the method in order to know whether this method has good results, one of them is by calculating the accuracy of the obtained classification results. In this research, applied confussion matrix method to get the result of its accuracy.

III. Result and Discussion

The method of extraction of HSI color features applied in this study, yielded 12 features. Judging from the extraction, the color features of the beef look slightly different from the pork. However, these extraction results are not directly classified, but are combined with feature histogram texture extraction resulting in 6 features of each image. So each image has 18 features to be used as a training set on the classification process.

18 features available are used for classification with two classes, namely pork and beef. In this research applied 3 classification method that is neural network, knn, and svm using tools RapidMiner. From the classification results, can be calculated the level of accuracy using confusion matrix, and obtained the following results:

TABLE I. Confussion Matrix of SVM

SVM	true beef	true pork	Precision
Predictionbeef	16	0	100.00%
Prediction pork	24	40	62.50%
Recall	40.00%	100.00%	
Accuracy = 70%			

TABLE II. Confussion Matrix of KNN

KNN	true beef	true pork	Precision
Predictionbeef	33	5	86.84%
Prediction pork	7	35	83.33%
Recall	82.50%	87.50%	
Accuracy = 85%			

TABLE III. Confussion Matrix of NN

NN	True beef	True pork	Precision
Predictionbeef	37	2	94.87%
Predictionpork	3	38	92.68%
Recall	92.50%	95.00%	
Accuracy = 93.75%			

IV. Conclusion

In this study, evaluation of each classification shows that SVM had 70% of its accuracy, KNN had 85%, and neural network results is the bestproduces an accuracy that is 93.75% with k-folds 20. So it can be concluded that combining features between HSI color features with histogram-based texture features can classify the image of meat with 2 labels, beef and pork well using Neural Network classification.

References

[1] K. Adi, S. Pujiyanto, O. D. Nurhayati, and A. Pamungkas, "Beef quality identification using color analysis and k-nearest neighbor classification," Proc. - 2015 4th Int. Conf. Instrumentation, Commun. Inf. Technol. Biomed. Eng. ICICI-BME 2015, pp. 180–184, 2016.

[2] K. Chen, X. Sun, and X. Tang, "Color Grading of Beef Fat by Using Computer Vision and Support Vector Machine", Computer and Electronics in Agriculture, pp. 27-32, 2010.

[3] M. E. Ali et al., "Analysis of pork adulteration in commercial meatballs targeting porcine-specific mitochondrial cytochrome b gene by TaqMan probe real-time polymerase chain reaction," Meat Sci., vol. 91, no. 4, pp. 454–459, 2012.

[4] L. Handayani, Jasril, E. Budianita, W. Oktista, R. Hadi, D. Fattah, R. Yendra, and A. Fudholi, " Comparison Of Target Probabilistic Neural Networ (PNN) Classification For Beef and Pork", Jornal of Theoritical and Applied Information Technology, Vol. 95, No. 12, 2017.

[5] Sachin D. Khirade, A.B. Patil, "Plant Disease Detection Using Image Processing", International Converence on Computing Communication Control and Automation, 2015.

[6] Moacir Ponti, Tiago S. Nazare, Gabriela S. Thume, “Image Quantizatiom as a Dimensionality Reduction Procedure in Color and Texture Feature Extraction”, 2015.

[7] Ricardus Anggi Pramunendar, Catur Supriyanto, Dwi Hermawan Novianto, Ignatus Ngesti Yuwono, Guruh Fajar Shidik, Pulung Nurtanto Andono, “A Classification Method of Coconut Wood Quality Based On Gray Level Co-Occurence Matrices”, International Converence on Robotic, Biomimetics, Intelligent Computational Systems (ROBIONETICS), pp. 25-27, 2013.

[8] L. Liu, M. O. Ngadi, S. O. Prasher, and C. Gariépy, “Categorization of pork quality using Gabor filter-based hyperspectral imaging technology,” J. Food Eng., vol. 99, no. 3, pp. 284–293, 2010.

[9] D. Laha and M. Hazarika, “ScienceDirect A heuristic approach based on Euclidean distance matrix for the machine-part cell formation problem,” Mater. Today Proc., vol. 4, no. 2, pp. 1442–1451, 2017.

[10] Mr. Hrishikesh P. Kanjalkar, Prof. S.S.Lokhande, “Feature Extraction of Leaf Disease”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)., vol. 3, 2014

[11] S. Kumar and H. Rajaguru, “Code Converters with City Block Distance Measures for Classifying Epilepsy from EEG Signals,” Procedia - Procedia Comput. Sci., vol. 87, pp. 5–11, 2016.

[12] A. I. Ropodi, D. E. Pavlidis, F. Mohareb, E. Z. Panagou, and G. J. E. Nychas, “Multispectral image analysis approach to detect adulteration of beef and pork in raw meats,” Food Res. Int., vol. 67, pp. 12–18, 2015.

[13] H. P. Kanjalkar and P. S. S. Lokhande, “Feature Extraction of Leaf Diseases,” vol. 3, no. 1, pp. 153–155, 2014.

[14] N. H. C. of E. R. Jaya, S. (Department of Computer Science and Engineering, New Horizon College of Engineering)Kumar, Mohan (Department of Computer Science and Engineering, “a Study on Data Mining Techniques , Methods , Tools and Applications in Various Industries,” vol. 8, no. 4, pp. 34–39, 2016.

[15] A. Pourreza, W. S. Lee, R. Ehsani, J. K. Schueller, and E. Raveh, “An optimum method for real-time in-field detection of Huanglongbing disease using a vision sensor,” *Comput. Electron. Agric.*, vol. 110, no. 1, pp. 221–232, 2015.

[16] J. Chairunisa. 'Dataset', 2018. [Online]. Available: https://docs.dinus.ac.id/index.php/s/zDIRHocJah05OaG. [Accessed: 4-April- 2018].

[17] R. Jaya, S , Mohan Kumar, “A Study on Data Mining Techniques, Mothods, Tool and Application In Various Industries”, Int J Cur Res Rev, vol. 8, 2016.

Gossip Algorithm Implementation for Network Protocol

Aditya Nugraha Tama[1] , Hartanto Kusuma Wardana[2] , Saptadi Nugroho[3]
Electronics and Computer Engineering Department
Satya Wacana Christian University
Salatiga, Indonesia
[1]martinusditya@gmail.com, [2]hartanto.kusuma@staff.uksw.edu, [3]saptadinugroho@gmail.com

***Abstract*— Gossip Algorithm is an algorithm which regulates the data exchange process between interconnected nodes. All of the nodes are configured to run the active and passive thread, and there is no node acting as a client or server. The active thread is initiated by choosing an IP address randomly. A node with the chosen IP address will be engaged to conduct the data exchange process. Whereas, the passive thread will wait until there is an invitation from another node to carry out the data exchange process. This work implements the Gossip algorithm into hardware such as Raspberry Pi which represents nodes. The connection between the nodes which exchange data can be represented in a form of graph. When there is a turned-off node, the other node connected to it can automatically detect the turn off and find a substitute node for the data exchange process. The same thing happens when there is a node which has just been active, the node will automatically find another node randomly to do the data exchange process. The result of the implementation of gossip algorithm is shown by a graph presenting that all of the nodes are interconnected to one another and the data contained in every node has the same value. The possibility of the formation of two separated graph can be overcome by adding more protection on the process of neighbor-election.**

Keywords—gossip algorithm, protocol, network, node

I. INTRODUCTION

The advance in computer and technology which has grown so rapid makes a lot of aspects in life become easier. Computer can help and even take over human works such as processing and analysing data, controlling and adjusting production processes, and connecting billions of other computers spreading all around the world through the internet network. Through internet, computers are able to exchange data by fulfilling a particular communication protocol which is available. In other words, computers are able to perform a communication process just like humans do.

In life, human beings will certainly carry out a communication process with others. One of the many forms of communication done by humans is gossiping. Gossip is one of the ways to communicate favored by a lot of people, as it has been done since the Roman times up until now. When human beings are gossiping, there is a process of information exchange, so that new specific information which is previously unknown or unheard may be obtained. Even though the information earned is not necessarily true, gossip may spread in a short period of time. Much like real-life rumor- and virus-spreading, information disseminated by a gossip protocol spreads quickly and reliably with high probability [1]. This phenomenon then triggers the scientists and scholars to enable computers to spread information as if they are human beings spreading gossips.

Gossip algorithm was originally founded when, there is a problem about how many calls are required where There are n ladies, and each of them knows some item of gossip not known to the others. They communicate by telephone, and whenever one lady calls another, they tell each other all that they know at that time [2]. Gossip algorithms are one of the most important communication primitives in large networks, and have been studied under di□erent names such as rumor spreading, information dissemination, and broadcasting [3].

Gossip algorithm consist of two thread, the active thread and passive thread. The active thread will try to send a request to random node, while the passive thread will always listen if there are any request from another node. In this work, gossip algorithm implemented in python language on Raspberry Pi. TCP/IP is used as a network protocol which ensure connection between nodes.

II. MATERIALS AND METHODS

In the system, all of the nodes are of equal position (there is no client or server). On each node, there are three essential things which are active thread, passive thread, and the data employed in the communication process. In order to be connected to each other, each node is connected to a router can be seen in Fig. 1.

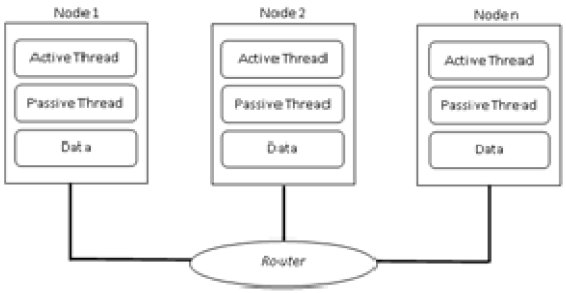


Fig. 1. Overview of the System

A communication graph $G = (V, E \subseteq V \times V)$ of n nodes is changing over time in terms of vertices and edges [4]. Each node will automatically run the passive thread, which is initiated by waiting for the submission from other nodes. Meanwhile, the initial node will run the active thread and choose an IP address to carry out the process of data exchange. After being chosen, the process of spreading gossips will begin. It will later on be forwarded by other nodes. Thus, in that particular system, all of the nodes agree towards a particular value related to a specific data. The generic protocol, described above, is a push-pull gossip protocol [5]. Rumors can be performed by transmitting a message in any round by any node along the edges in the graph [6].

A. Active Thread

Active Thread

1. **Wait** *for* **t** *seconds*
2. α <= *Select Random Neighbor*
3. *Send data to α*
4. **Wait until** *α send reply*
5. *Receive and Update data*

Fig. 2. Pseudocode active thread

In Fig. 2. Active thread is initiated by waiting for a predetermined time interval (t second). After that, the neighbor-choosing function is conducted. By utilizing the telnet and SSH facility, and by carrying out the *ip arp print* command, the list of IP address which is connected to the router can be found out. After the list of IP address is obtained, a random IP address is chosen. Then, the data will be sent out to the random IP. When there is a data reply from the neighbor, receive and accept the data, then count the average and update the data. If the neighbor does not reply, repeat from the beginning. In this system, the active thread works continuously (unlimited recurrence).

B. Passive Thread

Neighbor-choosing

1. **Login** *throught network gateway*
2. **If** *login successful:*
3. **Run** *ip arp print*
4. *Save the result*
5. *Filter the data*
6. α <= *Select Random Neighbor*
7. **End If**

Fig. 3. Pseudocode neighbor-choosing function

When passive thread is being run, the nodes will wait until there is another node inviting to do the data exchange. If there is a data submission received, the node will keep the data submission, then will send the data as a reply. Then, the average is counted and the data is updated. This process can be seen in Fig. 3.

Passive Thread

1. **Wait** *until receive data*
2. *Send data to sender*
3. *Receive and Update data*

Fig. 4. Pseudocode passive thread

C. Realization of the System

In order to realize the system, some tools are required. Those are Raspberry Pi as the node, router as the connector between networks and the proponent for the data communication process between nodes, and an access point as the connector of wireless network. Five Raspberry Pi are used as nodes. Three nodes are connected directly to the router, while the other two nodes are adjusted to a wireless network with the help of the access point as seen in Fig. 5. For the wireless network, the IP address is arranged dynamically by utilizing DHCP.

Prefix that be use in this network is /24. So that in one network there are 254 hosts that can be used. There are four networks used. All networks are in class C. IP address allocation shown in Table I.

TABLE I. IP ADDRESS ALOCATION IN THE SYSTEM

Name	Network ID	IP Address
Node 1	192.168.10.0/24	192.168.10.2
Node 2	192.168.20.0/24	192.168.20.2
Node 3	192.168.30.0/24	192.168.30.2
Node 4	192.168.60.0/24	192.168.60.xx (DHCP)
Node 5	192.168.60.0/24	192.168.60.xx (DHCP)

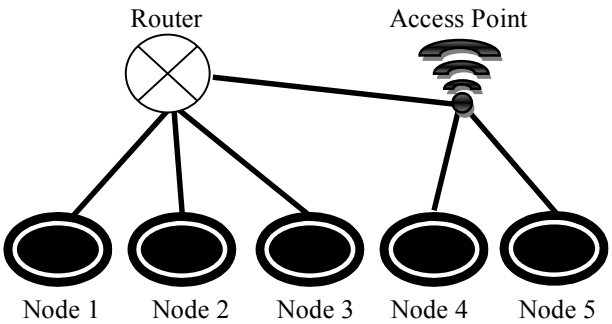


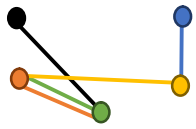
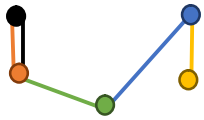
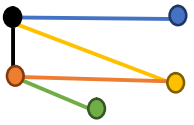
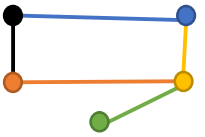
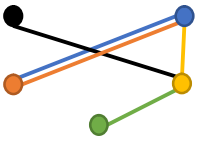
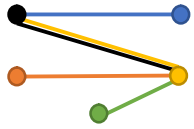
Fig. 5. A network showing a router, an access point and five Raspberry Pi that need to communicate

III. RESULT AND ANALYSIS

A. Graph Formation

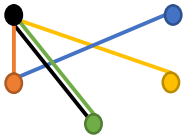
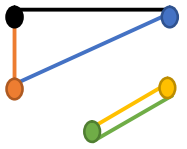
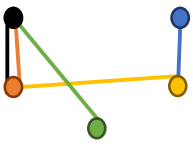
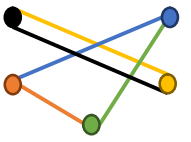
In this part, the testing is done 20 times, 10 times are for the testing of the node activation without protection for neighbor-choosing and 10 times for the testing of the node activation with protection for neighbor-choosing. The relationship between nodes is represented in a graph and the data communicated between nodes is recorded. The data that is used is determined as follows: the initial data for node 1 is 20, the initial data for node 2 is 25, the initial data for node 3 is 30, the initial data for node 4 is 35, and the initial data for node 5 is 40.

TABLE II. THE RESULT OF ACTIVATION NODES WITHOUT PROTECTION IN GETTING NEIGHBOR PROCESS

Result Num.	Node Activation Order	Graph	Initial Data	Data After Graph Exist
1	2=>5=>1 =>3=>4		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 30.95 2 = 30.95 3 = 30.95 4 = 30.95 5 = 30.95
2	3=>1=>5 =>4=>2		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 30.35 2 = 30.35 3 = 30.35 4 = 30.35 5 = 30.35
3	5=>1=>3 =>4=>2		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 30.83 2 = 30.83 3 = 30.83 4 = 30.83 5 = 30.83
4	2=>4=>1 =>5=>3		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 30.55 2 = 30.55 3 = 30.55 4 = 30.55 5 = 30.55
5	4=>1=>5 =>3=>2		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 30.28 2 = 30.28 3 = 30.28 4 = 30.28 5 = 30.28
6	2=>5=>3 =>4=>1		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 30.33 2 = 30.33 3 = 30.33 4 = 30.33 5 = 30.33

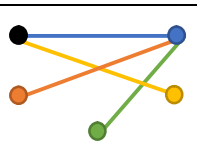
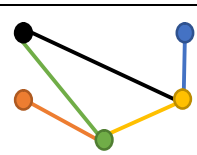
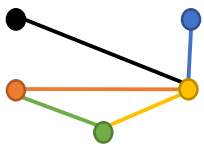
Information:
Node 1 Node 2 Node 3 Node 4 Node 5

TABLE II. THE RESULT OF ACTIVATION NODES WITHOUT PROTECTION IN GETTING NEIGHBOR PROCESS (CONTINUED)

Result Num.	Node Activation Order	Graph	Initial Data	Data After Graph Exist
7	3=>4=>5 =>1=>2		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 30.46 2 = 30.46 3 = 30.46 4 = 30.46 5 = 30.46
8	1=>4=>3 =>5=>2		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 25 2 = 25 3 = 25 4 = 37.5 5 = 37.5
9	5=>3=>1 =>4=>2		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 30.78 2 = 30.78 3 = 30.78 4 = 30.78 5 = 30.78
10	5=>3=>2 =>1=>4		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 27.5 2 = 33.75 3 = 33.75 4 = 27.5 5 = 33.75

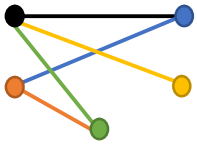
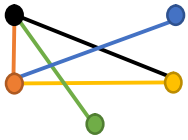
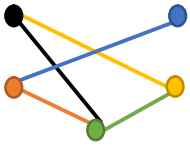
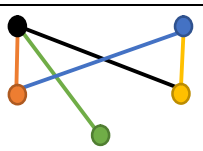
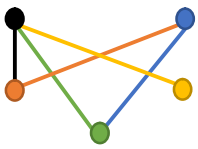
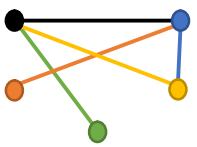
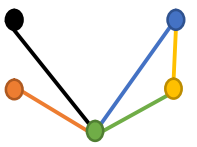
Information:
Node 1 Node 2 Node 3 Node 4 Node 5

TABLE III. THE RESULT OF ACTIVATION NODES WITH PROTECTION IN GETTING NEIGHBOR PROCESS

Result Num.	Node Activation Order	Graph	Initial Data	Data After Graph Exist
1	2=>5=>1 =>3=>4		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 26.3 2 = 26.3 3 = 26.3 4 = 26.3 5 = 26.3
2	3=>1=>5 =>4=>2		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 32.34 2 = 32.34 3 = 32.34 4 = 32.34 5 = 32.34
3	5=>1=>3 =>4=>2		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 29.68 2 = 29.68 3 = 29.68 4 = 29.68 5 = 29.68

Information:
Node 1 Node 2 Node 3 Node 4 Node 5

TABLE III. THE RESULT OF ACTIVATION NODES WITH PROTECTION IN GETTING NEIGHBOR PROCESS (CONTINUED)

Result Num.	Node Activation Order	Graph	Initial Data	Data After Graph Exist
4	2=>4=>1 =>5=>3		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 30 2 = 30 3 = 30 4 = 30 5 = 30
5	4=>1=>5 =>3=>2		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 28.06 2 = 28.06 3 = 28.06 4 = 28.06 5 = 28.06
6	2=>5=>3 =>4=>1		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 32.4 2 = 32.4 3 = 32.4 4 = 32.4 5 = 32.4
7	3=>4=>5 =>1=>2		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 30 2 = 30 3 = 30 4 = 30 5 = 30
8	1=>4=>3 =>5=>2		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 28.75 2 = 28.75 3 = 28.75 4 = 28.75 5 = 28.75
9	5=>3=>1 =>4=>2		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 29.99 2 = 29.99 3 = 29.99 4 = 29.99 5 = 29.99
10	5=>3=>2 =>1=>4		1 = 20 2 = 25 3 = 30 4 = 35 5 = 40	1 = 30.45 2 = 30.45 3 = 30.45 4 = 30.45 5 = 30.45

Information:
Node 1 Node 2 Node 3 Node 4 Node 5

The result of the testing shown in Table II. That in 8 out of 10 graph that are formed, there is a node which chooses to exchange data with node which has been connected to it, which is on the testing number 1, 2, 5, 6, 7, 8, 9, and 10. Even on the testing number 8 and 10, two separated graphs are formed, resulting in different amounts of value of the data on each node. This phenomenon becomes the drawback of gossip algorithm of random function without protection.

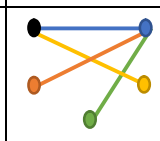
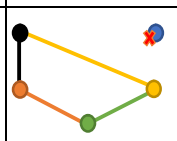
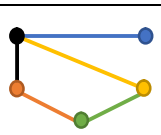
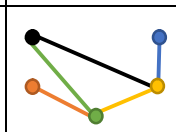
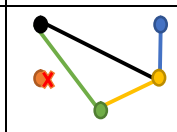
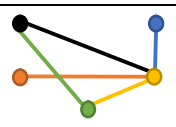
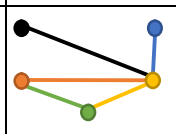
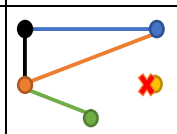
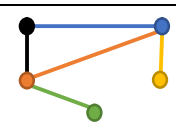
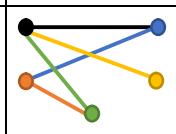
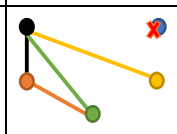
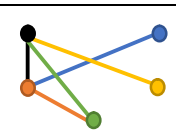
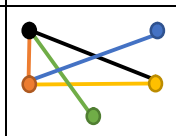
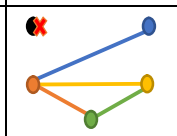
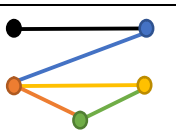
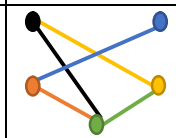
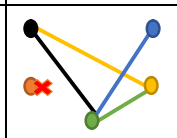
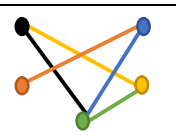
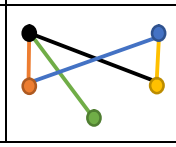
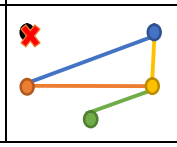
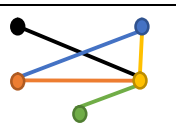
Table III is a table containing the result of the testing of the node activation with additional protection for neighbor-choosing. Of all the graphs formed, there is no node which reconnects a node which has already been connected to it. It proves that the additional protection in neighbor-choosing guarantees that all nodes are interconnected in a graph.

On both of the testing tables, the initial data and the data obtained after the graph is formed is written to find out whether the process of data delivery and updates have been conducted. On the testing the data can be communicated and updated according to gossip algorithm. There are indeed two conditions (testing number 8 and 10) on Table II where the data from the node has different values, or in other words, not all of the nodes agree upon a particular value. However, on each of the connected graph, the process of data communication and updates keeps on running according to gossip algorithm.

B. *Graph Transformation*

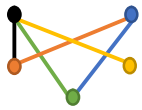
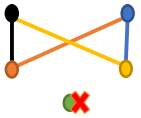
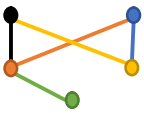
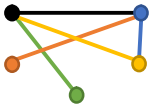
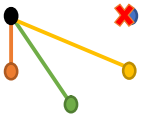
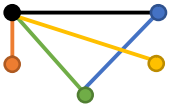
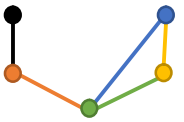
In the implementation which has been conducted, there are two things which influence the form of the graph. The first thing is when there is a turned-off node and the second is when there is a newly activated node.

TABLE IV. THE RESULT OF GRAPH TRANSFORMATION

Result Num.	Initial Graph	Graph After Node Go Off	Graph After Node Reactivate
1			
2			
3			
4			
5			
6			
7			

Information:
Node 1 Node 2 Node 3 Node 4 Node 5

TABLE IV. THE RESULT OF GRAPH TRANSFORMATION (CONTINUED)

Result Num.	Initial Graph	Graph After Node Go Off	Graph After Node Reactivate
8			
9			
10			

Information:

Node 1 Node 2 Node 3 Node 4 Node 5

From Table IV, it can be seen that when a node is turned off, the mechanism to find a substitute node to exchange data is successful. From the 10 testing, there happens a condition where a node cannot find a substitute node to exchange data, which is on testing number 9. It happens as a result of the protection for neighbor-choosing, because all of the active nodes have been connected to node 1 so there is no more node which can be chosen by node 1.

When there is a newly activated node, then the node will find a neighbor to exchange data. Of the 10 testing conducted on all of the testing a new node can connect to another node randomly.

IV. CONCLUSION

The process of neighbor-choosing with a risk of forming two graphs which may cause disconnection on nodes, where not all of the nodes are connected, can be overcome by adding a protection on the process of neighbor-choosing. The process of data delivery and updates between interconnected nodes has been successfully implemented. It is proven by the same data value when all of the nodes are connected. When there is a turned-off node, the node connected to it manages to choose a substitute node randomly. From the testing, there is a condition where a node has been connected to all of the active nodes, so when the node is performing the neighbor-choosing function, as a result of the protection for neighbor-choosing, there is no other node that is available. When there is a newly activated node, it can be connected to a node chosen randomly.

REFERENCES

[1] L. Alvisi, dkk, "How robust are gossip-based communication protocols?", ACM SIGOPS Operating Systems Review, pp. 14-18, 2007.

[2] Brenda Baker, Robert Shostak, "Gossip and Telephones", North Holland Publishing Company, pp. 191-193, 1971.

[3] Zeyu Guo, He Sun, "Gossip vs. Markov Chains, and Randomness-Efficient Rumor Spreading", SODA '15 Proceedings of the twenty-sixthannual ACM-SIAM symposium on Discrete algorithms, pp. 411-430, 2015.

[4] Devavrat Shah, "Gossip Algorithms", Foundations and Trends in Networking Vol. 3, No. 1 (2008) 1–125, 2009.

[5] Rena Bakhshi, Daniela Gavidia, dkk., "A modeling framework for gossip-based information spread", In Quantitative Evaluation of Systems (QEST) 2011 Eighth International Conference, pp. 245–254, IEEE, 2011.

[6] R. Karp, C. Schindelhauer, S. Shenker, B. Vocking, "Randomized Rumor Spreading", Proceedings 41st Annual Symposium on Foundations of Computer Science, IEEE, pp. 565-574, 2000.

A Robust Non-Blind Image Watermarking Method using 2-Level HWT-DCT

Ajib Susanto, De Rosal Ignatius Moses Setiadi, Eko Hari Rachmawanto, Christy Atika Sari
Department of Informatics Engineering, Dian Nuswantoro University
207 Imam Bonjol Street
Semarang 50131, Indonesia
e-mail: ajib.susanto@dsn.dinus.ac.id, moses@dsn.dinus.ac.id, eko.hari@dsn.dinus.ac.id, atika.sari@dsn.dinus.ac.id

***Abstract*— The current technology allows to distribute digital images quickly and easily over the internet. Some digital images have copyright that needs to be protected. Watermarking is a technique for protecting copyright. In order for copyright is not easily lost with image manipulation, then watermarking technique must be robust. This study proposes a combination of Discrete Cosine Transform (DCT) based on two levels of the Haar Wavelet Transform (HWT) on image watermarking method with the aim of improving aspects of robustness and imperceptibility. The first step, host image is transformed with DWT as much as two levels to get LL2 subband. LL2 subband is a subband containing the image core information of two levels of wavelet transformation. Furthermore, LL2 subband split into small sub-blocks with the size of 4 * 4. The sub-blocks are transformed again using DCT. Insertion of watermarks using alpha variables with a certain value on the DC coefficient matrix. This value is used for managing the quality and strength of the watermark. To test the imperceptibility and watermark resistance used measuring instruments MSE, PSNR, and correlation coefficient. Based on the simulation results obtained a watermarked image is quite satisfactory with a relatively good imperceptibility value and resistant to various attacks.**

Keywords— Copyright Protection, Haar, Image Watermarking, Robust, Wavelet

I. INTRODUCTION

The current technology allows distributing digital data quickly and easily over the internet. Some digital data have copyright that needs to be protected. Watermarking is a technique that is often used for copyright protection of digital data such as images, audio, and video. Along with the development of technology, the tools to manipulate digital data is also more sophisticated, especially digital images. This is the background why research on digital image watermarking continues to be developed. There are three types of watermarking according to its durability [1], ie: robust, fragile and semi-fragile. Robust Watermarking is used as a tool for copyright protection because of its resistance to manipulation. Fragile is very susceptible to manipulation even with a bit of manipulation making the watermarks disappear, so this method is suitable for authentication. While the semi-fragile just hold up against certain attacks only.

Some important properties in watermarking are fidelity, robustness, payload data and blind or non-blind detection [2]. Fidelity is a quality perception of watermarks, this means the harder the watermark detected the human senses the better the quality of fidelity. Robustness is resistance to various attacks. While blind or non-blind means the need to use a cover image for the extraction process or not.

There are various watermarking techniques developed, such as Least Significant Bit (LSB), Fourier Transform, Cosine Transform, and Wavelet Transform [3] [4] [5]. LSB is included in the spatial domain, ie watermark insertion techniques that directly manipulate the cover image. While the above transformation is a technique that is included in the frequency domain. Image watermarking done in the frequency domain is more resistant to various attacks [6].

DWT is a domain transformation that has good multiresolution characters in human vision systems and also works well in spatial localization [7]. DWT has also been used as a standard transformation of JPEG 2000 [8]. DCT is an orthogonal transformation that transforms images into two coefficients ie Alternating Current (AC) and Direct Current (DC). The computational complexity of DCT is relatively simpler than DWT [6]. DCT also fits in real-time imagery, since DCT has been used as a standard JPEG compression has been widely embedded on various hardware [8]. So by combining these two transformations can be obtained more optimal results.

Haar Wavelet Transform (HWT) is a DWT given a haar filter, with the addition of a haar filter, data security and watermark capacity can be improved [9]. In a previous study conducted by Susanto et al [8], it was proposed a combination of HWT and DCT to optimize watermark insertion. The research results proved by the combination of both transformations can improve the fidelity and robustness aspects. In the study [10] the use of two DWT levels combined with DCT was shown to improve both aspects. This study proposed a combination of two levels HWT and DCT on a Non-Blind method to improve both aspects.

II. THE STATE OF THE ART

Research on watermarking in digital images is not a new thing. There has been much research on published watermarking images. Many researchers continue to innovate to find better watermarking techniques. Aspects aimed at image watermarking research is robustness, imperceptibility, and payload. While this research focuses on aspects of two aspects, namely robustness, and imperceptibility. Suasato et al [8], proposed a combination of HWT and DCT transformations for watermarking techniques on grayscale imagery. HWT 1-level transformation is done first with DCT based on subblock with size 8 * 8. This technique produces a 43dB PSNR on Lena's image and has proven resistant to various image manipulation attacks.

Another study conducted by Winarno et al [6], proposed a combination of wavelet and DCT transformation. The difference of DCT transform is done first based on subblock with size 8 * 8. The result of transformation with DCT produces DC coefficient which then collected for wavelet transformation. Subband LL is selected to insert a watermark. Imperceptibility test results with PSNR yielded 42,69dB on Lena grayscale image. This method is also resistant to various attacks, especially JPEG compression.

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

Sudibyo et al [9], also proposed an HWT transformation to apply watermarking techniques. In his research, the HWT transformation is combined with the Chinese Remainder Theorem (CRT) algorithm to encrypt the watermark. The watermark insertion technique uses the blind method so it does not require cover image during extraction. The results of the proposed method have a high degree of imperceptibility, with a PSNR value of about 63dB. But this technique is only strong against Gaussian noise attacks.

Another study conducted by Arya et al [7] also combines DCT and DWT techniques. The proposed technique is quite similar to the technique used by Susanto et al [8], the difference being the use of Haar filters on wavelet transforms and insertion techniques that do not use alpha values but use addition and subtraction values based on the watermark pixel value. The result of this method gets the PSNR value reaching 51dB and is strong enough against various attack manipulations.

The combination of DWT and DCT is also proposed on watermarking techniques in Ali's research [10], also combining DWT and DCT. In addition, also applied the PN-Sequence to insert a watermark. DWT transformation is done 2-level followed by DCT. Extraction method used is a blind method. The study also compared the DWT transformation alone with the combination of DWT-DCT. The test results show that the DWT-DCT combination gets better results in imperceptibility and robustness aspects. From some research related above, hence in this research proposes a combination method of HWT-DCT in the non-blind method. The 2-level HWT transformation is done first to get the LL2 subband, followed by DCT transformation on LL2 to insert the watermark.

III. TRANSFORMATION DOMAIN ON IMAGE WATERMARKING

A. Haar Wavelet Transform (HWT)

DWT is one of the domain transformations that is widely implemented in watermarking technique because it has many advantages, one of which is its superior character in decomposing the image [4]. In the DWT buffer has a variety of filters such as two filters are widely used Daubechies and Haar. The use of haar filters on DWT is often referred to as HWT. The haar filter has advantages in computational speed and is done recursively [8]. The wavelet transform uses two filters to decompose the image. The filter is a low pass filter and a high pass filter. The filter is operated on every column and every image line to produce four sub-bands. Subband that contains image approximation is called LL. The subband containing the horizontal downsample value of the image is called HL. Subband that contains vertical downsample is called LH. The last subband is called HH, this subband contains the image diagonal downsample [9]. Watermark insertion can be done on one subband or some subband. In order for watermark resistant to manipulation then subband LL better chance. Because the watermark can fit in the image approximation value. It's just that by modifying the LL subband too much can curb the fidelity aspect. Fidelity aspect can be measured with PSNR unit, and if the PSNR value reaches at least 40dB then fidelity quality is included in the very good category. However, in the study [6] [7] [8], insertion of LL subband succeeds well with a value above 40dB. In this study, the insertion is proposed on subband LL2 by decomposition with two levels of DWT.

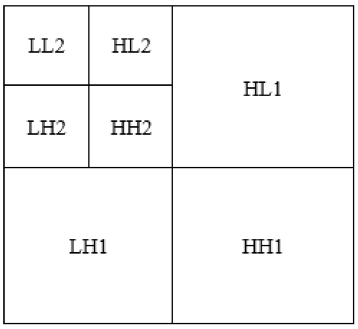


Fig. 1. Two Level Wavelet Transform pada subband LL

B. Discrete Cosine Transform (DCT)

DCT is a widely used domain transformation in image watermarking like DWT. Using DCT in multiple image watermarking schemes can make watermarks more resistant to JPEG compression manipulations [6]. This is certainly very important in copyright protection because many data submission applications that perform automatic compression on the image. DCT has an advantage in computing speed and energy concentration in the image when transformed. DCT will produce two kinds of coefficients, namely DC and AC [8]. The image core energy is stored on the DC coefficient. Other information is stored on AC coefficient. DC coefficients are found in the first column and the first line of transformation. By inserting a watermark in the DC section, the watermark will be harder to remove.

DC	AC	AC	AC
AC	AC	AC	AC
AC	AC	AC	AC
AC	AC	AC	AC

Fig. 2. DCT Transformation Results on 4x4 sub-block

IV. PROPOSED METHOD

There are two main schemes in watermarking techniques. The two schemes are watermark embedding schemes and watermark extraction schemes. Extraction scheme uses a non-blind method where required cover image. The following details of the insertion and extraction scheme are proposed in the proposed scheme.

A. Proposed Embbeding Scheme

The embedding scheme takes input that is the host image and watermark image. While the output of the embedding process produces a watermarked image. Here are the details of the proposed embedding scheme.

Step 1 : Read the host image, then do the transformation with DWT, resulting LL, HL, LH, and HH subbands.

Step 2 : Select the LL subbands, then do the transform again with DWT so as to generate LL2, HL2, LH2, and HH2 subbands.

Step 3 : Select subband LL2, then break the subband into small sub-blocks of size 4 * 4.

Step 4 : Perform transformations on each sub-block with DCT transformation, then generated AC and DC coefficients.

Step 5 : Select the DC coefficient to place the watermark with a specified alpha value (α). Note, alpha values can affect aspects of imperceptibility and robustness in the watermark image.

Step 6 : Once pinned on DC coefficients do inverse DCT, and two levels of DWT to generate a watermarked image.

To see a more detailed picture of the watermark embedding process, see figure 3.

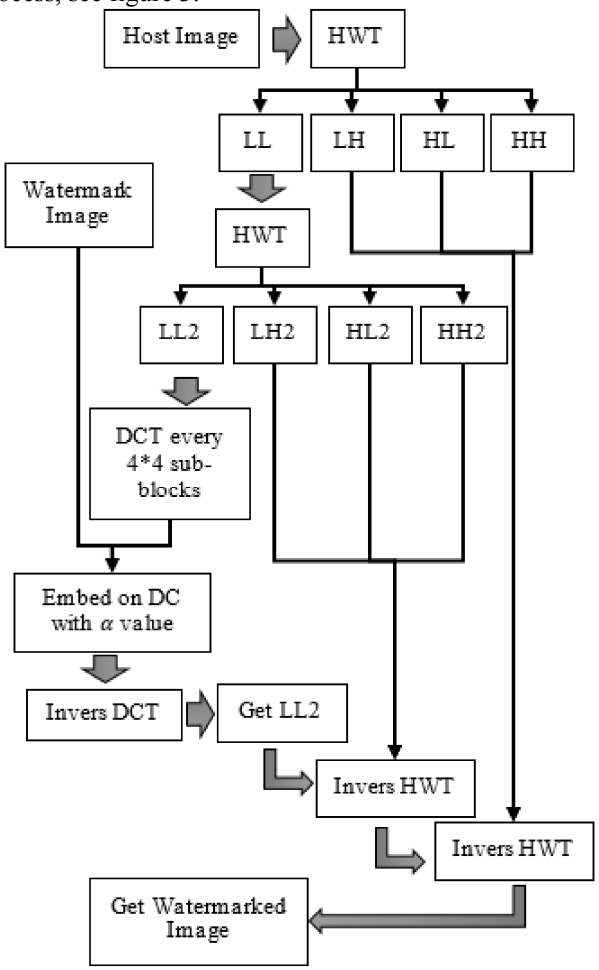


Fig. 3. Proposed Embbeding Scheme

B. Proposed Extraction Scheme

The extraction scheme proposed in this research is a non-blind method, so it takes input that is the image of the host and the watermarked image. Both images will be compared so as to produce watermark image extraction. Here are the details of the proposed extraction scheme.

Step 1 : Read the image host and watermarked image, then do the transformation of both images with DWT, thus producing four kinds of a subband (LL, HL, LH, and HH).

Step 2 : Select subband LL, then do the transform again with DWT so as to generate subband (LL2, HL2, LH2, and HH2).

Step 3 : Select subband LL2, then break the subband into small sub-blocks of size 4 * 4.

Step 4 : Perform transformations on each sub-block with DCT transformation, then produce AC and DC coefficients.

Step 5 : Select the DC coefficient for watermark extraction by comparing the DC value of the host and the watermark using the specified alpha value (α) in the embedding process.

Step 6 : Without the inverse transformation get the watermark image of the extraction process.

To see a more detailed description of the watermark embedding process, see figure 4.

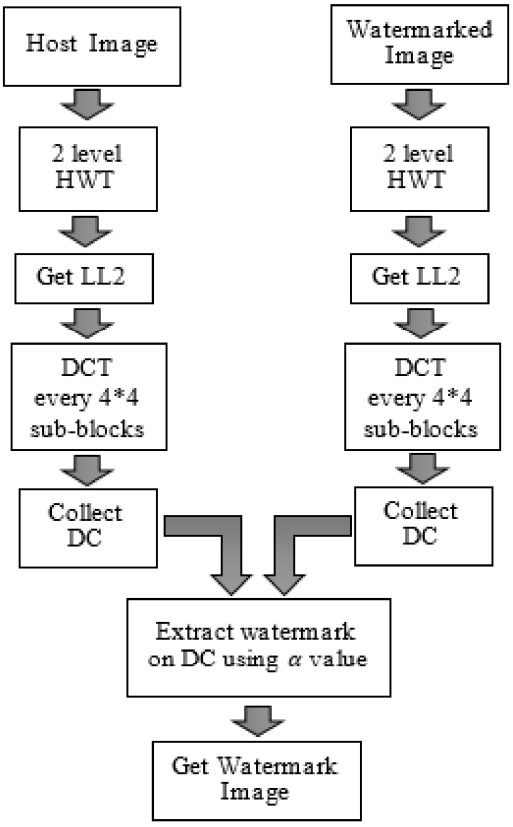
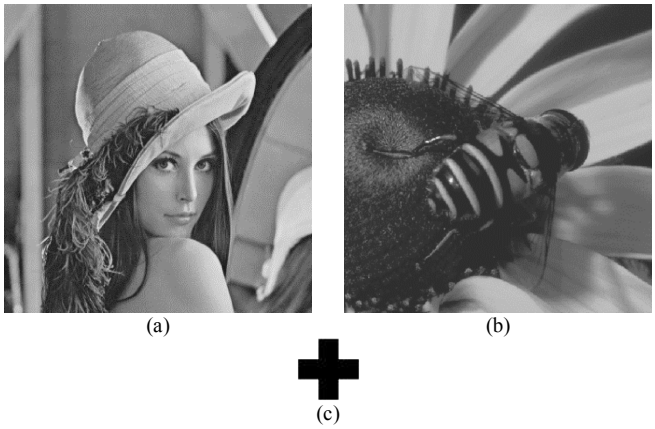


Fig. 4. Proposed Extraction Scheme

V. SIMULATION AND TESTING

This section will perform a simulation to evaluate the performance of the proposed method. In the simulation stage is a grayscale image with size 512 * 512 used as a host image, while watermark image is a binary image with size 32 * 32. Figure 5 shows the picture used in this research.



(a) (b)

(c)

Fig. 5. (a)-(b) Host image used {(a)Lena; (b) bee; } (c) Watermark image

Aspects to be measured performance, namely aspects of fidelity and robustness. The fidelity aspect will be measured with PSNR and MSE [11]. MSE can be calculated by (1), while PSNR is calculated by (2) [12].

$$MSE = \sum_{m=0}^{M} \sum_{n=0}^{N} \|h(m,n) - w(m,n)\| \quad (1)$$

$$PSNR = 10\ log_{10}\left(\frac{255^2}{\sqrt{MSE}}\right) \quad (2)$$

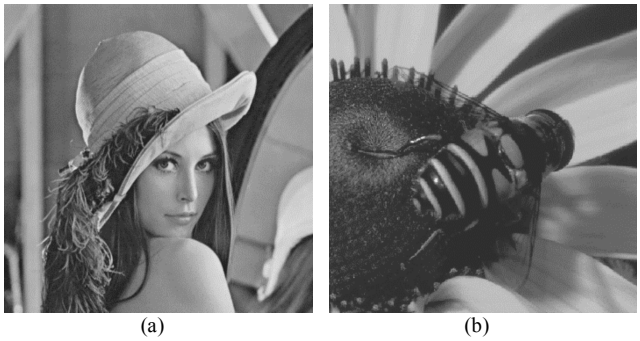
Where:

m, n = the number of rows and columns

h = the host image

w = the watermarked image

Based on the measurement result, both images get the value of PSNR 45.0199 dB and MSE value 2,0469. The value may still be in either category because the PSNR value is above 40 dB. This result is also better than the previous study, where with Lena host image and the same watermark obtained only PSNR value 43,221 dB [8] These results are also better than the proposed method in [6], to see the watermarked image of the insertion process can be seen in figure 6.



(a) (b)

Fig. 6. (a)watermarked Lena; (b) watermarked bee

The second simulation is to test aspects of watermark resistance to various attacks. In this study, there are eight kinds of attacks tested. The attack will be performed on the host image, then the watermark is extracted from the host image that has been attacked. The extraction results will be measured by comparing the original watermark correlation with the extracted watermark. The formula used for correlation calculation is the correlation coefficient (3). The correlation coefficient is a widely used method for bathing image correlation. The range value is between 0 and 1, where the value 0 is very bad or has no similarity at all. While the value of 1 is a sign that the extraction results can be done perfectly.

$$cc = \frac{\sum_m \sum_n (w_{mn} - \overline{w})(w'_{mn} - \overline{w'})}{\sqrt{(\sum_m \sum_n (w_{mn} - \overline{w})\ (\sum_m \sum_n R_{mn} - \overline{w'})}} \quad (3)$$

Where w' is extracted the watermak image.

Based on the measurement of the correlation coefficient on the extraction results, all obtained value 1. This indicates that the extraction process can be done perfectly. This is very important to know before the attack test. Because if without attack the value of correlation is not good, then in theory with the provision of attacks on the image host will produce a bad correlation. Table 1 shows the results of watermark extraction after an attack on the host image.

TABLE I. CORRELATION COEFFICIENT RESULTS FROM PROPOSED METHOD

Attack	Lena Image	Bee Image
No Attack	 1.0000	 1.0000
JPEG compression	 0.9134	 0.9213
Median filter	 0.9605	 0.8371
Unsharp filter	 0.8779	 0.9062
Salt & pepper noise	 0.8075	 0.7523
Gaussian noise	 0.6285	 0.6711
Blur	 0.8766	 0.9250
Low Pass Filter	 0.8548	 0.7979
Cropping	 0.8874	 0.8874

VI. CONCLUSIONS

This study proposes a combination of two transformations, namely DWT and DCT, to improve results, both in terms of quality of their imperceptibility and robustness. Based on the results of simulations and tests that have been done can be concluded that the proposed method can work well. The use of two DWT levels proved to improve the imperceptibility aspect compared to one DWT level. The method of extraction also works well where the results of the extraction without attack get the value 1. Watermark also proved to survive the eight kinds of attacks that have been tested. Especially in JPEG compression attacks, this method proved reliable with the highest average CC value, that is 0.9176.

REFERENCES

[1] A. C. Bovik, The Essential Guide to Image Processing, Austin: Academic Press, 2009.

[2] I. J. Cox, M. L. Miller, J. A. Bloom, J. Fridrich and T. Kalker, Digital Watermarking and Steganography, 2nd ed., E. Inc, Ed., Morgan Kauffman, 2008.

[3] O. Abodena, M. Agoyi, and E. Celebi, "Hybrid Technique for Robust Image Watermarking Using Discrete Time Fourier Transform," in *Signal Processing and Communications Applications Conference (SIU)*, Antalya, 2017.

[4] W. S. Sari, E. H. Rachmawanto, D. R. I. M. Setiadi and C. A. Sari, "A Good Performance OTP Encryption Image based on DCT-DWT Steganography," *TELKOMNIKA Telecommunication, Computing, Electronics and Control,* vol. 15, no. 4, pp. 1987-1995, 2017.

[5] E. Elbasi and V. Kaya, "Robust Medical Image Watermarking Using Frequency Domain and Least Significant Bits Algorithms," in

International Conference on Computing Sciences and Engineering (ICCSE), Kuwait, 2018.

[6] A. Winarno, D. R. I. M. Setiadi, A. A. Arrasyid, C. A. Sari, and E. H. Rachmawanto, "Image Watermarking using Low Wavelet Subband based on 8×8 Sub-block DCT," in *International Seminar on Application for Technology of Information and Communication (ISemantic)*, Semarang, 2017.

[7] R. K. Arya, S. Singh, and R. Saharan, "A Secure Non-blind Block Based Digital Image Watermarking Technique using DWT and DCT," in *International Conference on Advances in Computing, Communications, and Informatics (ICACCI)*, Kochi, 2015.

[8] A. Susanto, D. R. I. M. Setiadi, C. A. Sari and E. H. Rachmawanto, "Hybrid Method using HWT-DCT for Image Watermarking," in *International Conference on Information Technology for Cyber and IT Service Management (CITSM)*, Denpasar, 2017.

[9] U. Sudibyo, F. Eranisa, E. H. Rachmawanto, D. R. I. M. Setiadi and C. A. Sari, "A Secure Image Watermarking using Chinese Remainder Theorem Based on Haar Wavelet Transform," in *International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, Semarang, 2017.

10] A. Al-Haj, "Combined DWT-DCT Digital Image Watermarking," *Journal of Computer Science,* vol. 3, pp. 740-746, September 2007.

[11] S. D. Daphal, P. K. Aher, S. B. Galande and S. K. Jagtap, "Comparison of key transform domain watermarking methods based on performance measures," in *International Conference on Communication information and Computing Technology (ICCICT)*, Mumbai, 2018.

[12] D. R. I. M. Setiadi and J. Jumanto, "An Enhanced LSB-Image Steganography Using the Hybrid Canny-Sobel Edge Detection," *Cybernetics and Information Technologies,* vol. 18, no. 2, pp. 74-88, 2018.

Classification Method of Academic Acquisition Value Through Student Behavior

Nur Cahyo Nugroho, Aditya Galih Setyaji, Guruh Fajar Shidik
Magister Teknik Informatika, Universitas Dian Nuswantoro
Email: mas.nugroho@gmail.com, galih.research@gmail.com,
guruh.fajar@research.dinus.ac.id

Abstract—Educational data mining becomes inseparable part in research and development. Abundance of data that related to education is very useful to help education development in the world. One kind of data that can be extract from education learning, student achievements or behavioural studies. Variety of data that are currently available in field of education makes variation of performance measurement research students attributes. Student behavior are highly related to psychological information on learning achievements and took action against predictions. Data mining using behavior attribute category showed significant influence on achievement. Algorithm like J48, Naive Bayes, MLP and SMO preceded by a normalization on preprocessing obtained the highest accuracy results 79.16 on Neural Network base algorithm with Multi Layer Perceptron. SVM base algorithm SMO behaviour attribute category using descritze on pre-processing shown best accuracy on 66,062% to predicting acquisition value.

Keyword—*Educational Data Mining, behavioral student, e- learning, discretization, normalize, SMO, MLP*

1 Introduction

A Wide range of educational data we can get at this time. Educational data can be used in decision support to educational institution development. Online system that exist at each educational institution increasingly facilitates the retrieval of the data periodically. Data source can be obtained with a variety of both from traditional way with direct observation of teaching based on experience, a survey and interviews, or from student database have been recorded from information system in educational institutions.

Through the behaviour and learning activities of students at the online record during the process of learning e-learning system. Behavioral attributes for performance measurement of students is still unpopular. During this time, performance student assessment using derived data from values obtained, personal data, or questionnaire or interview from associated students of learning activity. Data mining is very potential big data type information from the e-learning learning by knowing information behavior through activity users on real time system [1].

Selection of diverse research attributes will greatly influence results of study even though using same method. EDM used to make predictions, cluster analysis, classification, association rule etc. [1]. In a review, Muhammed and Tasir (2013) deliver a platform e-learning model used make prediction, classification, clustering etc to analyze learning model.

Famous algorithm frequently used among others Decission Tree, Naive Bayes, NN, SVM, KNN, DBSCAN and Expectation Maximization Clustering (EMC) [2]. In supervised learning we have trainee data and consist set of training examples. This paperfocus on supervised learning.

This paper is divided into five sections. Research that has been conducted about the prediction of performance assessment of students described in section II of the review of the literature. Section III describes the research methodology at this writing. After that in section IV will discusses the results of research and discussion in section V was final result of the conclusions of this paper.

2 Related Research

Educational Data Mining is widely used in decision-making system of the most widely used for classification, clustering or prediction. Research on existing algorithms used to extract student performance data, among others, decision tree, neural network, naive bayes, K-Nearest Neighbours, logistic regression and Support Vector Machine. (SVM) [3]. Variety of data that currently available in field of education makes variation of performance measurement research students attributes.

EDM used to probe possibility of failure of learning students use Decission Tree with Naive Bayes, J48, SVM and Neural Network [4]. On this research aimed to evaluate against the EDM technique at the time used to make early predictions of failure students. Through two techniques sample data retrieval education through distance learning (e-learning) and learning in classroom yield different results when carried out on preprocessing against source of data.

Georgios Kostopoulos et al [5] research on distance learning using Active Learning. Attributes that use demographic, academic background and value test method using prediction done J48, JRip, Multi Layer representation of Perceptron NN, NB and SMO that represent SVM. Results obtained illustrate performance of SMO is higher than other four methods.

Popular methods are used in EDM technique will differ greatly and depend on the performance of the category attribute [2]. Mohammed Shahiri did a review use method of Neural Network, Decision Tree, Naive Bayes, k-NN and

SVM. With different attributes, result accuracy using Decision Tree 66% - 90%, Neural Network 69% - 98%, Naive Bayes 50% - 65%, k Nearest Neighbours 69% - 83% and SVM 80 - 83% [3].

Gigle and Zhai (2017) using e-learning instructtional data by method of Markov Model for predicting the performance of students. Using attribute behavior students doing behavioral differences based on identification of clusters [6]. Development methods of collecting data began to shift in the educational model of distance learning will provide extensive research and variation amount large data.

Research on student behavior conducted earlier by Amrieh et al[7] shows a strong influence between student behavior and achievement of values. The classification used in this study are Decission Tree, Naive Bayes and Neural Network (NN) then compare student behavior attributes and without behavioral attributes. results of several classification methods obtained accuracy of 29% between the use of behavioral attribute categories and without behavioral attribute categories.

This article is focus on evaluation supervised learning such as Decission Tree, Naive Bayes, SVM and NN to compare of each algorithm on prediction with a single category attribute in behavior or student interaction.

3 METHODOLOGY

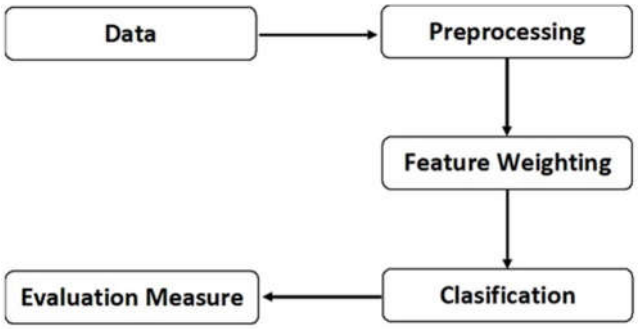


Figure 1. Methodology

3.1 *Data*

In this research, the author uses dataset repository from kaggle.com obtained from the e-learning system or commonly called Kalboard 360. Through this e-learning system retrieved the dataset that contains the attribute table behavior, background demographics, survey and background education as well as the label in the form of the final value of graduation. Experiments data obtained student activity from tracking tool or students commonly referred to experience API (X-API) in the Kalboard System.

Dataset in this study is the public data that is in the kaggle.com that have been handed out beginning in 2016. Collecting method using the e-learning system Experience API that does event tracking student behavior in learning as read an article, view announcement and watch video tutorial. Data e- learning has been through preprocessing before data cleaning against the total amount of data as much as 500 sample. Experiment data as much as 480 instances and consists of 16 attributes that are categorized into demographics, educational background, parent participant and student behavior.

Data experiment used has been made preprocessing in earlier research Amrieh et al (2015) with discretization in category label into three intervals, where low level (L) has a value of 0 - 69, mid-level (M) has a value of 70 - 89 and high level (H) has a value of 90 - 100 [7]. Whole process in this experiment using WEKA to do preprocessing, feature weighting and classification model comparison. In addition, this experiment used a 10 fold cross validation in conducting training and testing data.

Table I
STUDENT ATTRIBUTE

Category	Attribute
Demographics	Nationality
	Gender
	Place of Birth
	Relation
Educational Background	Stage
	Grade
	Semester
	Topics
	Section
Parents Participation	Survey Answering
	Parent Satisfaction
Student Behavior	Raised Hand
	Open the Reference
	Viewing Announcement
	Absence
	Discussion Group

3.2 *Preprocessing*

Preprocessing is a process which took place prior to the next algorithms process. This process aims to improve performance algorithm used. This research will compare of preprocessing used either normalize and discretization on behavioral attribute. Normalize is a preprocessing by changing artibut has a value of integer real number into integer between Min = 0 and Max = 1.

3.3 *Feature Weighting*

Feature weighting is one things that are important to look a level of influence of each atributes against label. At this stage the author using the method of Information Gain (IG) and the method of Rank to see the magnitude of influence attribute [2].

Table II
ATTRIBUTE WEIGHTING

Attribute	Weights	Attribute	Weights
Nationality	0.1277	Topics	0.0761
Gender	0.0517	Absence	0.3974
Place of Birth	0.1229	Survey Ans	0.1504
Relation	0.1261	Satisfaction	0.1067
Stage	0.01058	Raised Hand	0.3733
Attribute	Weights	Attribute	Weights
Grade	0.04748	Open the reference	0.458
Section	0.00703	View Announc.	0.2578

Semester	0.01182	Group Discussion	0.1139

3.4 *Classification*

Analysis aims attribute gives the effect on acquisition value as seen from student behavior during instruction. Classification Method used in research is Decission tree (using the J48), Naive Bayes, NN (using Multi Layer Perceptron) and SVM (using SMO). Algorithm is populer method use in Educational Data Mining.

1. Decission Tree (J48) [1], [2], [3]
   J48 algorithm is an implementation of C4.5 on tree-based WEKA. J48 uses entropy and information gain for determination of root early. Entropy is parameter for measuring heterogenity of research data, which ever more heterogeneous data with larger value entropy
2. Naive Bayes [1], [2], [3]
   Naive Bayes is a learning method using probabilitas as a basic condition. This method is derived from Bayes theorem.

$$P(C|X) = \frac{P(X|C)\ x\ P(C)}{P(X)} \quad (1)$$

3. Multi Layer Perceptron [8]
   Classification of Multi layer neural network which is perceptron has a number of neurons connected by weights and output [5].
4. Sequential Minimal Optimization (SMO)
   SMO is an improvement algorithm of Support Vector Machines (SVMs) used to perform the analysis by performing optimization function [9].

3.5 *Evaluation Measure*

Evaluation measure look the quality of the classification use Accuracy, Precision, Recall and F-Measure [10][11]. The measurement can be calculated based on equation 2, 3, 4 and 5 below:

Table III
CONFUSION MATRIX

		Prediction	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Posifif (FP)	True Negative (TN)

1. Accuracy is a measurement model results and proportion of the amount data are classified correctly and total data [2]. TP measurement is describes the ”True Positive”, FP illustrates ”False Positive”, FN describes ”False Negative” and TN is described ”True Negative”.

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (2)$$

2. Precission referred as Positive Predictive Value (PPV) which is in ratio of correct classification with a total classification wrong and right [2]. Precission described with formula:

$$Precission = \frac{TP}{TP+FP} \quad (3)$$

3. Recall is the ratio of correct classification with a total of which is not classified correctly and correct classification[2]. Recall is described as follows.

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

4. F-measure is a measurement using combination of precision and recall to bring up relationship indicators[2].

$$F - Measure = 2x\ \frac{Precission\ x\ Recall}{Precission + Recall} \quad (5)$$

4 RESULT AND DISCUSSION

Experiment research using WEKA software program 3.8.1 for evaluation method. Further, in con- ducting on training and testing using 10 fold cross Validation and performed a comparison results obtained from each method. This discussion focuses on process of comparison preprocessing using four algoritm method. Preprocessing analysis are using discretization and normalize in behavior attribute Raised Hand, Open the Reference, Viewing Announcement and Group Discussion.

Experiment using classification algorithm proposed shows SMO gained highest accuracy rate of 78.75. Root Mean Square Error (RMSE) classification results 0.3522 and Mean Square Error (MAE) 0.3522.

Table IV
CLASSIFICATION WITHOUT PREPROCESSING

Measurements Evaluation	J48	Naive Bayes	MLP	SMO
Accuracy	75.833	67.708	78.333	78.75
Precision	0.76	0.675	0.783	0.788
Recall	0.758	0.677	0.783	0.788
F-Measure	0.759	0.671	0.783	0.787

Table V
CLASSIFICATION WITH NORMALIZE

Evaluation Measurements	J48	Naive Bayes	MLP	SMO
Accuracy	75.833	67.708	79.166	78.75
Precision	0.76	0.675	0.791	0.788
Recall	0.758	0.677	0.792	0.788
F-Measure	0.759	0.671	0.791	0.787

Table VI
CLASSIFICATION WITH DESCRETIZATION

Measurements Evaluation	J48	Naive Bayes	MLP	SMO
Accuracy	71.875	73.75	78.542	76.042
Precision	0.718	0.736	0.787	0.761
Recall	0.719	0.738	0.785	0.76
F-Measure	0.718	0.735	0.784	0.761

Experiment result using normalize showed improved accuracy of Multi Layer perseptron from 78.333 became 79.166 with RSME 0.3448 and MAE 0.1485 but other methods that not different result.

Other comparison results classification with and without preprosesing, next experiment using standalone category behavior attribute without other categories. This experiment will explore accuracy of behaviour category for acquicition prediction .

Table VII
ATTRIBUTE BEHAVIOR WITHOUT PREPROCESSING

Evaluation Measurements	J48	Naive Bayes	MLP	SMO
Accuracy	65	62.917	62.708	61.25
Precision	0.66	0.632	0.628	0.61
Recall	0.65	0.629	0.627	0.613
F-Measure	0.646	0.616	0.624	0.606

Table VIII
BEHAVIOR WITH NORMALIZE

Evaluation Measurements	J48	Naive Bayes	MLP	SMO
Accuracy	65	62.917	62.708	61.25
Precision	0.66	0.632	0.628	0.61
Recall	0.65	0.629	0.627	0.613
F-Measure	0.646	0.616	0.624	0.606

Table IX
BEHAVIOUR WITH DISCRETIZATION

Evaluation Measurements	J48	Naive Bayes	MLP	SMO
Accuracy	64.583	64.375	62.917	66.042
Precision	0.648	0.642	0.63	0.684
Recall	0.646	0.644	0.629	0.66
F-Measure	0.644	0.64	0.627	0.652

Experiment results using classification category attribute behavior there is no difference in results of calculations without preprocessing and normalize. Preprocessing with discretice improve accuracy on SMO algorithm quite significant reach 5% from 61.25% to 66.042% with RSME 0.396 and MAE 0.305. Experiment result on influence of behavioral attributes is pretty well to predict acquisition value using SMO algorithm.

5 CONCLUSION

Behavior attributes are generally significant influence to predictions on education dataset. Experiments results using behavior attribute categories, demographics, educational background and satisfaction showing Neural Network had the highest accuracy among other algorithms. After a preprocessing using normalize able to improve the performance of neural network algorithm.

In standalone behavior category attribute also good enough to predict value acqusition on SVM base algoritm using SMO. With using discretice on preprocessing improve accuracy using the SMO algorithm quite significant reach 5% from 61.25% to 66.042% with RSME 0.396 and MAE 0.305.

Category behavior attribute individually is also well used to make predictions and have an influence higher than category attribute. Four attribute Raised Hand, Open the reference, See the Announcement and Group Discussion in this experiment can improve using another behaviour attribute category. This research is still allowed for development used attribute task assignment and GPA on each semester is includes for comprehensive analysis.

REFERENCES

[1] S. K. Mohamad and Z. Tasir, "Educational Data Mining: A Review," *Procedia - Soc. Behav. Sci.*, vol. 97, pp. 320–324, 2013.

[2] B. Mehboob, R. M. Liaqat, and N. A. Saqib, "Predicting Student Performance and Risk Analysis by Using Data Mining Approach," *Int. J. Comput. Sci. Inf. Secur.*, vol. 14, no. 7, pp. 69–76, 2016.

[3] A. Mohamed, W. Husain, and A. Rashid, "The Third Information Systems International Conference A Review on Predicting Student ' s Performance using Data Mining Techniques," *Procedia - Procedia Comput. Sci.*, vol. 72, pp. 414–422, 2015.

[4] E. B. Costa, B. Fonseca, M. A. Santana, F. F. de Araújo, and J. Rego, "Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses," *Comput. Human Behav.*, vol. 73, pp. 247–256, 2017.

[5] G. Kostopoulos and A. Lipitakis, "Predicting Student Performance in Distance Higher Education Using Active Learning," vol. 1, no. Dm, pp. 75–86.

[6] C. Geigle, "Modeling MOOC Student Behavior With Two-Layer Hidden Markov Models," vol. 9, no. 1, pp. 1–24, 2017.

[7] E. A. Amrieh and I. Aljarah, "Preprocessing and Analyzing Educational Data Set Using X- API for Improving Student ' s Performance," 2015.

[8] M. Tinchcombe, "Multilayer Feedforward Networks are Universal Approximators," vol. 2, pp. 359–366, 1989.

[9] J. C. Platt, "Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines," *Adv. kernel methods*, pp. 185–208, 1998.

[10] T. Y. Chen and R. Merkel, "On the Statistical Properties of the F-measure," pp. 0–7.

[11] D. M. W. POWERS, "Evaluation: From Precision, Recall and F-Measure To Roc, Informedness, Markedness & Correlation," *J. Mach. Learn. Technol.*, vol. 2, no. 1, pp. 37–63, 2011.

A Comparative Study of some Video Players based on Visual Output Quality

Christy Atika Sari, Hasda Nuril Mala, Eko Hari Rachmawanto, De Rosal Ignatius Moses Setiadi, Desi Purwanti Kusumaningrum
Department of Informatics Engineering
Faculty of Computer Science, Dian Nuswantoro University
Semarang, Indonesia
E-mail : atika.sari@dsn.dinus.ac.id, 111201408536@mhs.dinus.ac.id, eko.hari@dsn.dinus.ac.id, moses@dsn.dinus.ac.id, desi.purwanti@dsn.dinus.ac.id

***Abstract*— Video is a media that is often used by humans for entertainment or learning. The video player application is used to play a video. In fact, many video player apps are in circulation today. Each of these applications has different qualities, both in terms of visual and in terms of audio. This study was conducted to determine the visual quality generated by frequently used video player applications such as KMP, GOM, MPC, and VLC. To measure the visual quality of the video can use the Structural Similarity Index Measure (SSIM). SSIM was chosen because it is more accurate, consistent and correlated with human visuals. The research stages begin with the process of video acquisition, video framing, capturing, and measurement. The results of this study show that KMP applications are able to provide better visual quality, where the application produces the highest overall SSIM average value of 0.99778. This result is an objective assessment that can be a recommendation for choosing a video player application.**

Keywords—Visual Quality; Video Player; Capturing; SSIM; Image Quality

I. INTRODUCTION

The video is one digital media that is growing rapidly in recent years and are used for various things, such as entertainment, learning, delivery of information, and communication. A video can be played using a video player. Technological advancements bring a variety of video player applications with different visual and audio quality [1]. This is due to various factors such as the standard format used by each developer and the uniqueness of the video player application. Applications like MPC-HC, KMPlayer, GOM, and VLC are pretty popular and much-installed video player apps [2]. The video is a medium that displays audio and visual simultaneously. Visual components that exist in the video is actually a collection of frames or digital image.

In digital image processing, image quality can be measured using several measuring tools such as Mean Squared Error(MSE), Peak Signal to Noise Ratio(PSNR), Structural Similarity Index Measure (SSIM), and others [3] [4] [5] [6]. In this study SSIM will be used to measure the visual quality of the video player application because SSIM is built or formed to replace the previous measuring tool This is because MSE and PSNR get great criticism because the measurement results obtained are not correlated well with the perspective of the human eye, SSIM which method is more consistent and more accurate [7] [8]. SSIM is a measuring instrument that uses component structures such as luminance (luminance / l), average (mean / μ), variant (σ^2), and co-variant (σ_xy) in digital images. This SSIM measuring tool is often used in several studies such as on [9] [10] [11] to measure the visual quality of images and video. The study was conducted to compare the visual quality generated by some video player applications by comparing the frames captured through those applications with the original frames of the downloaded videos obtained from the video acquisition process through the computation method. Each frame is a digital image. The visual capacity of the digital image will be measured with SSIM. The results of this study are expected to be useful to educate and provide an overview to select video player applications based on visual quality.

II. RESEARCH METHOD

There are four stages in the method that perform in the study, which is shown in Figure 1:

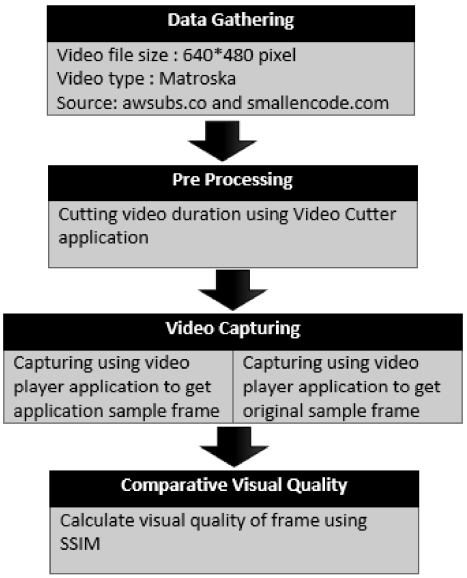


Fig. 1. Research Method

A. Data Gathering

The video file is an input in the process of this research. The file is a Matroska video file with a size of 640x480 pixels. Matroska video is a file with extension .mkv. Matroska is an open standard multimedia format that can hold many video files, audio, imagery, and subtitle-based WebM files [12]. This makes Matroska video files popular for video file storage. The video file used consists of two types of video, which is an animated video downloaded from awsubs.co and a live-action video downloaded from smallencode.com. The first video file is part of the Episode 2 movie Grandblue Fantasy The Animation (GFTA), while the second video file is part of the Ruler Master of The Mask (RMTM) Episode 17 movie. Some video player software is also collected at this step, such as MPC-HC version 1.7.8.61 (3eaa44a) (develop), KMP version 4.1.5.8 KMP Plus, GOM version 2.3.9.5265, and VLC version 2.2.4 Weatherwax. This four software will be measured with SSIM and compared which is better in terms of visual quality.

B. Preprocessing

This stage is done by cutting video movie duration. It is intended that the video file can be read by the Matlab application. The original movie duration is relatively large, which is about 27 minutes to make the video file cannot be read by Matlab, then the duration is trimmed to 47 seconds. The process of cutting video duration using Free Video Cutter software version 4.0.1. Figure 2 shows the preprocessing process in this study.

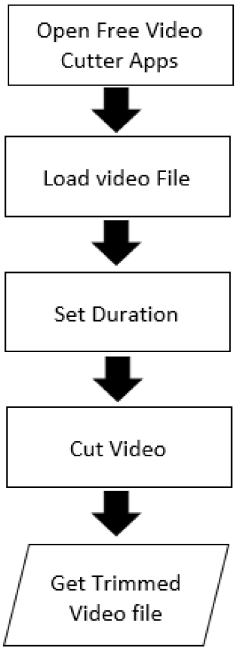


Fig. 2. Preprocessing steps

C. Capturing Frames from Video file

Video capturing process is the process of breaking the video into multiple frames. This video splitting uses two methods: a computational method using the Matlab application to get an original frame of video and capturing method by using video player application whose visual video result is tested to get test frame. The frame will be saved in the form of a digital image with PNG format. The PNG format is selected because this format does not remove information when compressed (lossless) [13]. Stages of capturing process conducted in this study can be seen in Figure 3.

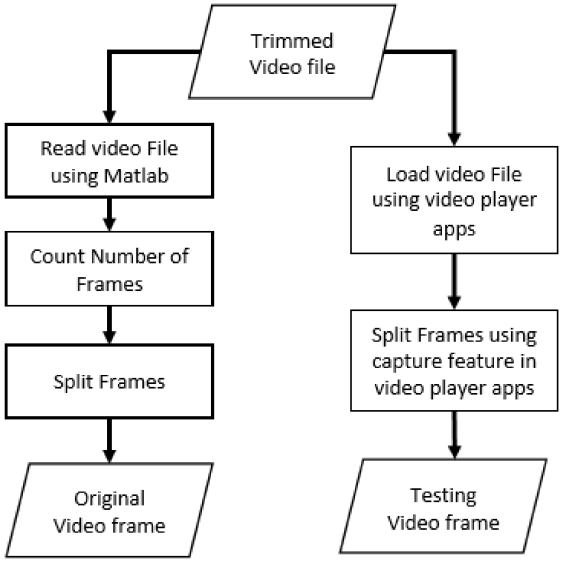


Fig. 3. Capturing video

D. Visual quality measurement using SSIM

The process of measuring the visual frame quality using SSIM, which is used to obtain the objective value of each frame generated by each video player application. Mathematically this SSIM measuring instrument can be formulated in (1).

$$SSIM = l_{x,y}^{\alpha} \times c_{x,y}^{\beta} \times s_{x,y}^{\gamma} \quad (1)$$

Where l represents lumination, c is the contrast, s is the structure of the frame, and α, β, and γ are the exponents indicating the importance of the three components (l, c, and s), where the value of the three exponents is> 0. The three components in the SSIM can calculate by the formula (2), (3), (4).

$$l_{x,y} = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (2)$$

$$c_{x,y} = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (3)$$

$$s_{x,y} = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (4)$$

Where μ_x is average of the test frame, μ_y is average of the original frmae, σ_x is covariance of test frame, σ_y isi covariance of original the frame, σ_{xy} is covariance test frame against original frthe ame, σ_x^2 is variant of test frame, σ_y^2 is variant of oria ginal frame, and c_1, c_2, dan c_3 are the values of constants

obtained from $c_1 = (k_1 L)^2$; $c_2 = (k_2 L)^2$; $c_3 = \frac{1}{2} c_2$, with the value L = 255 for the 8-bit image, is the frame being observed (test frame) and y is the original or actual frame. When the exponent value = 1, then the above SSIM formula can be simplified into formula (5).

$$SSIM = \frac{(2\mu_x\mu_y + C_1) \times (2\sigma_{x,y} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1) \times (\sigma_x^2 + \sigma_y^2 + C_2)} \quad (5)$$

III. RESULTS AND ANALYSIS

In preprocessing the results obtained in the form of the first video duration of more than 24 minutes and 30 minutes cut into video files that lasted 45 and 49 seconds. This process is done because of limited memory capacity in Matlab to read video file. The capturing process itself generates multiple frames, Figure 4 shows the sample frame of the GFTA movie, while Figure 5 shows the sample frame of the RMTM movie.

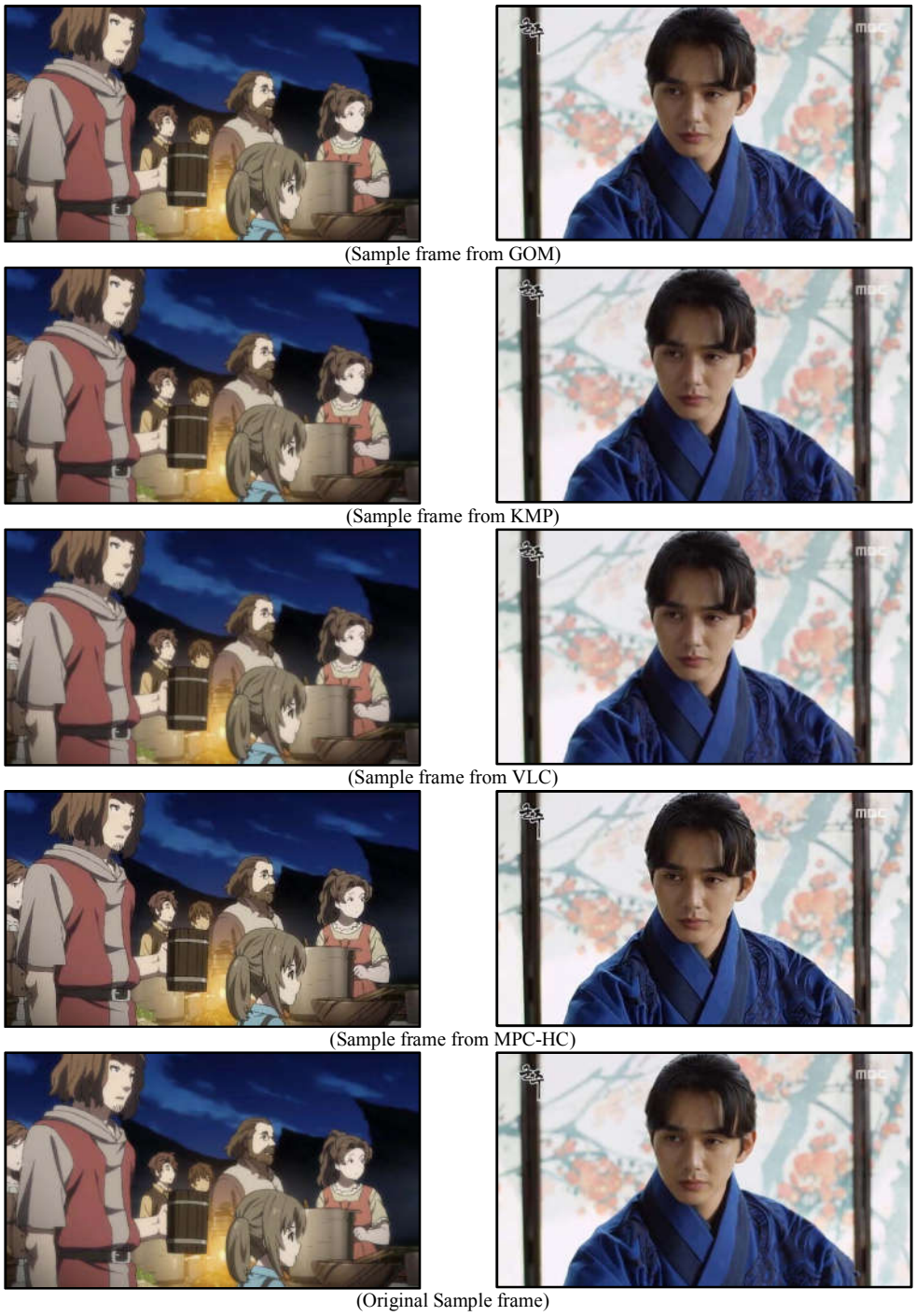


Fig. 4. Sample frame from GFTA and RMTM Movie

From the resulting frames, a measurement process using SSIM will be performed, Figure 5 is a diagram showing the SSIM values generated from the GFTA Movie frame

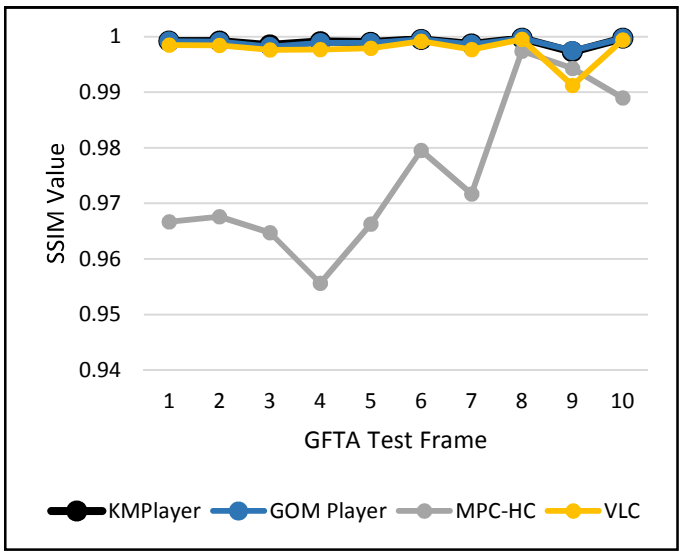


Fig. 5. Sample frame from GFTA

Based on the Figure 5 above measurements of SSIM measurements, it can be seen that the values generated by the test frames in GOM, KMP and VLC applications have almost the same quality level, whereas in MPC-HC applications have lower quality, where the results obtained have a distance which is quite distant from the value results in KMP, GOM and VLC except on the MPC test frame 8 to 10.

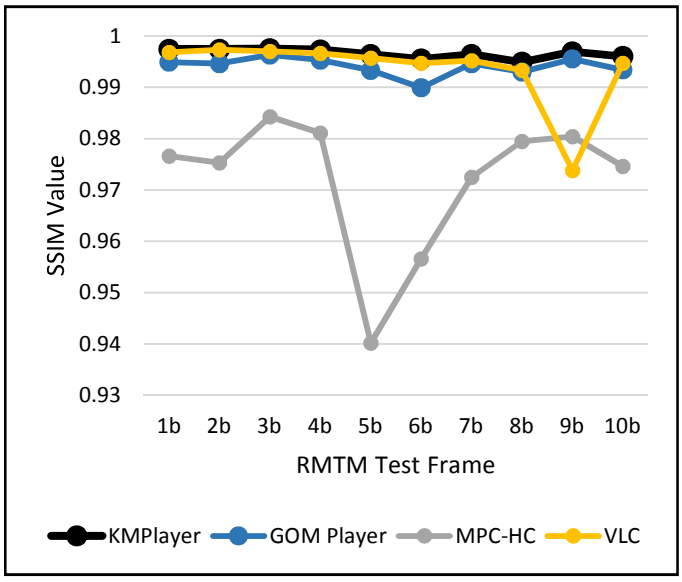


Fig. 6. Sample frame from GFTA

Figure 6 is a diagram of the SSIM values in the RMTM test frame. Based on Figure 6 it appears that KMP, GOM, and VLC applications also have almost the same quality. While the resulting test frame, MPC has less consistent values. Table 1 shows the average value of SSIM values obtained by each application.

TABLE I. AVERAGE SSIM FROM ALL FRAME

Average SSIM	KMP	GOM	MPC-HC	VLC
GFTA	0.99899	0.99893	0.97528	0.99771
RMTM	0.99657	0.99408	0.97211	0.99351
Average	0.99778	0.996505	0.973695	0.99561

From table 1 it can be concluded that the highest average SSIM value in both movies is generated by KMP application where the value indicates that the average value of the frame has similarity level which is close to the original frames.

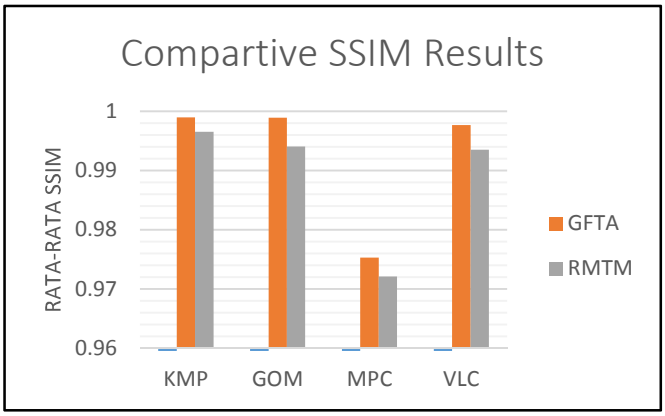


Fig. 7. Comparative SSIM Results

Based on Figure 7 it can be concluded that the visual quality produced by the application of the GOM, KMP and VLC does not change significantly as it did in the application MPC, as the original frames of course the quality it has a quality frame that is perfect so as to choose the quality of the frame is a test of the best is a frame with a quality that is very close to the original quality KMP is a video player application that produces the best visual video quality with a mean value of SSIM for GFTA is 0.99899 and for RMTM is 0.99657.

IV. CONCLUSIONS

The results of this study indicate that the application of KMP, GOM, and VLC has an average value of SSIM is closer to the value of 1, while the MPC application produces a lower average value compared to the three applications. Based on the results of this study can be concluded KMPlayer application is the best video player application that can be used by the community as a media video player. This is evidenced by the average value of SSIM obtained where the overall average value of KMPlayer is 0.99778.

REFERENCES

[1] M. B. Tirta and G. F. Shidik, "Evaluation performance of cloud computing with network attached storage for video render," in *International Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarang, 2017.

[2] U. Kumar, "Best DVD player apps for Windows 10 users," Windows report, 8 4 2016. [Online]. Available: https://windowsreport.com/best-dvd-players-windows-10/. [Accessed 2018].

[3] P. Cika, D. Kovac, and J. Bilek, "Software for Video Quality Assessment," in *International Conference on Telecommunications and Signal Processing (TSP)*, Vienna, 2016.

[4] D. R. I. M. Setiadi, H. A. Santoso, E. H. Rachmawanto and C. A. Sari, "An Improved Message Capacity and Security using Divide and Modulus Function in Spatial Domain Steganography," in *International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, 2018.

[5] D. R. I. M. Setiadi and J. Jumanto, "An Enhanced LSB-Image Steganography Using the Hybrid Canny-Sobel Edge Detection," *Cybernetics and Information Technologies,* vol. 18, no. 2, pp. 74-88, 2018.

[6] W. S. Sari, E. H. Rachmawanto, D. R. I. M. Setiadi and C. A. Sari, "A Good Performance OTP Encryption Image based on DCT-DWT Steganography," *TELKOMNIKA Telecommunication, Computing, Electronics and Control,* vol. 15, no. 4, pp. 1987-1995, 2017.

[7] P. W. Adi, F. Z. Rahmanti and L. Umaroh, "Improving CRT based Watermarking using Integer Wavelet Projection," in *International Conference on Innovative and Creative Information Technology (ICITech)*, Salatiga, 2017.

[8] P. Gurav dan G. Patil, "Full-Reference Video Quality Assessment using Structural Similarity (SSIM) Index," *MAT Journals,* vol. 1, no. 2, pp. 1-15, 2016.

[9] L. Lin, X. Liao, G. Tan, H. Jin, X. Yang, W. Zhang dan B. Li, "LiveRender: A Cloud Gaming System Based on Compressed Graphics Streaming," *IEEE/ACM Transactions on Networking ,* vol. 24, no. 4, pp. 2128 - 2139, November 2016.

[10] Y. A. Mekarsari, D. R. I. M. Setiadi, C. A. Sari, E. H. Rachmawanto and Muljono, "Non-Blind RGB Image Watermarking Technique using 2-Level Discrete Wavelet Transform and Singular Value Decomposition," in *International Conference on Information and Communications Technology (ICOIACT)*, Yogyakarta, 2018.

[11] M. L. Genco, T. Adaş, and N. Özbek, "Stereo Video Quality Assessment using SSIM and depth maps," in *Signal Processing and Communication Application Conference (SIU)*, Zonguldak, 2016.

[12] N. Pitropakis, C. Lambrinoudakis, D. Geneiatakis and D. Gritzalis, "A Practical Steganographic Approach for Matroska Based High Quality Video Files," in *International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, Barcelona, 2013.

[13] Z. Liu, X. Cheng, C. Jia and J. Yang, "Format Compliant Degradation for PNG Image," in *International Conference on Emerging Intelligent Data and Web Technologies (EIDWT)*, Xi'an, 2013.

Classification of Music Moods Based on CNN

Fika Ulfa Widowati
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
Fikaulfawidowati@gmail.com

Fitriyanto Santi Nugroho
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
elanggeni@gmail.com

Guruh Fajar Shidik
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
guruh.fajar@research.dinus.ac.id

Abstract*—In recent years, music has become part of everyone life. Almost everyone with different backgrounds listening to music, as a form of entertainment or just accompany in doing the activity. Music can be enjoyed in various ways based on the level of society. Usually music is categorized by genre, which shows the type of music or the form of music. Like pop, rock, metal, ballad, dangdut, jazz, keroncong, and etc. But the division by genre is common. So, in this study we try to classify a music into some moods like happy, peaceful, angry, and sad. Because such a classification will be easily accepted by listeners, especially people who are not very familiar with the genre. This classification is performed with 200 track databases for the training data and 50 track database for testing. We use a music features, such as pitch, pulse clarity, tempo, key, and scale. For the classification method in this paper we used Convolutional Neural Networks (CNN), the accuracy that we have are 82%.

Keywords— Classification; Data Mining; Moods; Music Information Retrieval; CNN

I. INTRODUCTION

Today music has become a very important part in human life. And almost everyone listens to music every time, and every time they do their activities. According to Miller [1], listening to favourites music for 30 minutes can help release endorphin hormones to relieve pain and reduce stress. Various music is said to have influence in their research, including one of them is classical music. Meanwhile, endorphin itself is a substance that causes a person to feel happy, happy, and active activities. While in research conducted by Ferguson [2], states that people can quickly improve their mood within 2 weeks just by listening to music. Both studies show how influential music is to our lives, and one of them is the influence of music on the mood of the listener. Therefore, in this paper we are discussing about music and mood. Usually music is categorized by genre, which shows the type of music or the form of music. Such as pop, rock, metal, dangdut, jazz, keroncong, dance, blues, hip hop, and etc. Although such categorization may help us in classifying music, it becomes ineffective because the listener likely to listen to music that suits his or her feeling (mood). Mood is become a complicated indicator compared to the genre, which is more common. Mood becomes more personal and very subjective because the song is considered sad can be different if heard by others. According to Stuart Laraia [4], the mood is an elongated emotional state that affects the entire personality of the individual and his life function. It is related to emotions and has the same understanding with the state of feeling/emotion. The conclusion is, mood becomes more personal and very subjective because the song is considered sad can be different if heard by others. And can also a song or music can give two or more different effects related to his mood. Then to simplify the research, in this project we choose four kinds of distinct moods, which Russell has proposed, that is happy, sad, angry, and peaceful [3] [5].

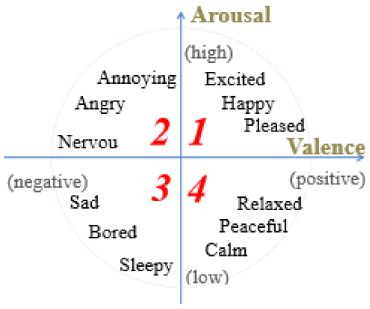


Fig. 1. The Russells valence-arousal emotion space [5].

The data training used for comparison is 200 songs, stored in the database [12]. The song has been classified based on arousal (excitement) and chemistry values by the song database makers. Two hundred songs taken for each mood used. For the data test, we used public dataset from www.kaggle.com. Datasets are labelled manually, i.e. (1) happy, (2) peaceful (3) sad, and (4) angry. The music features used in this project are pitch, pulse clarity, tempo, key, and scale. Fast music tempo used to detect the mood of music [9]. Many music features can be used in this research, but we choose the 5 most easily understood music features. While tools that we used are

MathLab 7.0. Then we use a classification algorithm to predict the label for a song. And then, for the training set we use 50 songs for each mood. So there are 200 songs in his training set. The classification algorithms we use for this project is Convolutional Neural Networks (CNN). CNN show very strong performance on practical sentence classification tasks (Kim, Kalchbrenner, Johnson (2014)), therefore we try to use CNN for different types of data processing, so we will know whether the accuracy will be more increased or not compared to the other classifications such as SVM, Naive Bayes, and Decision Tree that used in other studies [6] to classified the music.

II. Related Work

Thanh and Shirai, on their project use a lyrics and metadata that exist in a song to be classified with the Naive Bayes, SVM algorithms and graphical methods based on mood. Because that is used only writing metadata, so it cannot be used on the sound of a musical instrument.

Avijeet Kartikay, Harish Ganesan, Vanadana M Ladwani in his research used music using the Naive Bayes, Support Vector Machine (SVM) method, Linear Discriminant Analysis and Decision Trees with aim to classify music based on moods. The accuracy resulting from this classification were 59.8% 75, 4% and 71, 4%.

Lu, Liu, and Zhang on their research, state that western music is only determined from aspects of dataset, rhythm intensity, and timbre. Gaussian Mixture Model were used in the classification, while the moods are measured is depression, contentment, exuberance, and anxious.

Vallabha Hampiholi on his research use RMS Energy, Brightness, Low Energy Frames, Bandwidth, Roll-Off, Fluctuation, Zero Cross for use as a classification feature. Data used 122. Model classification using C45. With the mood label used is Exuberance Anxiety, Serene, and Depression. The accuracy resulting from this classification were 60% for Bollywood songs and 40% for Western songs.

Research conducted by Yang, Lin, Su, and Chen used a regression approach to recognize music by emotion. The goal is to predict the value of excitement and chemistry (AV) music track. Then, used Principal Component Analysis to reduce the relationship between 2 parameters, and choose the most appropriate feature to use. SVM is used for repressor.

III. Methodology



Fig. 2. Methodology.

A. Database

A data training of 744 songs used as a dataset, is the work of song database by Mohammad Soleymani, who sampled his song from FMA (Free Music Archive). The song which is a 45-second song. Each of the songs has been labelled with an excitement value and an average chemistry of the value provided by Amazon Mechanical Turk (AMT).

- Excitement : listener feels when heard to the song.

- Chemistry : the connection between the listener and the song.

The database stored in data-song folder, in a song info.csv and a static anno.csv which used to extract mean excitement and mean chemistry values that are need to prepare for our dataset [11]. All the data training that we used obtained from the internet [7].

B. Musical Feature

The remaining 5 musical features used as attributes in this study is based on research conducted by Oliver [8], which is as:

1. Pitch : Perceptual property of sound that allows their ordering on a frequency-related scale (Hz).
2. Pulse Clarity : How a listener can identify the rhythmic pulsation.
3. Tempo : The speed that the music should be played. It is measured in beats per minute (BPM).
4. Key : The set of pitches, or scale used in a song.
5. Scale : The pattern of note; major or minor scale.

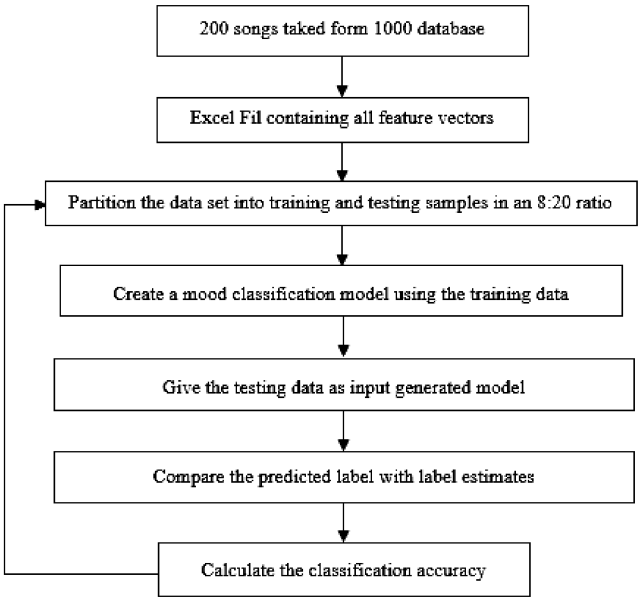


Fig. 3. Classification of Music.

C. Convolutional Neural Network Classification

The early development of the Convolutional Neural Network (CNN) can be traced from the [Fukushima 1980] designing an unmanned artificial network called Neocognitron, inspired by the Biological Visual Systems designed by the invention of Hubel and Wiesel-Hubel and Wiesel have won the Nobel Prize in physiology in 1981. LeCun et al. 1989 extends the model on all sides and applies it to supervised learning

contexts to solve handwriting recognition problems. The term " Convolution Neural Networks" it's self-introduced in LeCun et al. 1998. In its development, CNN save combined with Recurrent Nerve Networks (RNNs) that are often used for sequential models data such as audio signals or word order. This merger. This model is called Convolutional Recurrent Neural Network (CRNN).

In this study the concept that will be made is for the Convolutional Neural Network are:

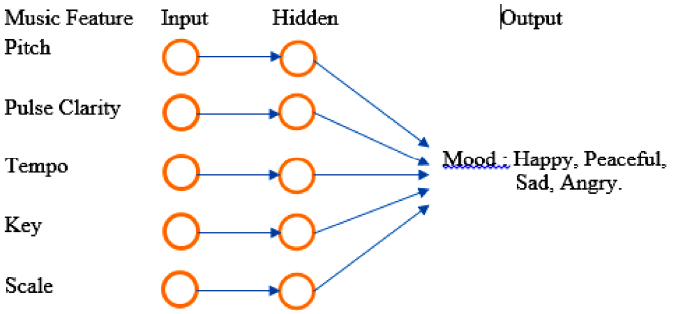


Fig. 4. Convolutional Neural Network.

CNN is perfect to maximize multinomial the logistic regression purpose of correct labelling. This is tantamount to corroborating the probability potential of L being defined by the number of log-probabilities of the right label throughout the training dataset D under the parameterized model by θ, represents as

$$L(D,\theta) = \sum_{i=1}^{|D|} \log\left(P(Y = y^{(i)}|x^{(i)},\theta)\right) \quad (1)$$

Where y(i) and x(i) = class and input unit corresponding to the i−th data sample, respectively. The prediction distributing is defining with the software function as

$$P(Y = i|x,\theta) = \frac{\exp(h_i)}{\sum_{j=1}^{C} \exp(h_j)} \quad (2)$$

Where hi and C = total input value to the output value i and show the number of classes[10].

D. Process

Convolutional Neural Network classification requires input features and output targets. The input feature is specifying from the sample value of the input song to training data by counting its music features. Once we have determined the classification of training data, we then simulate the classification. Then will get the division of mood. We have taken 200 songs (50 for each category). Then we have calculated the values of five features described above using MatLab. After that we get a neural layer that consist of 2 hidden layer and 5 neurons and 1 output layer with 3 neurons. the next step is train the network. network here should be set to reduce errors when predicting. After that we will test the data. from there we can see whether the results are in accordance with what we want.

IV. DISCUSSION

Five songs feature audio option stored in one feature. This feature is used as a feature input matrix with 5 rows of value features and 200 columns for 200 tracks, to train artificial neural networks. Each of these columns represents a feature of a single song. The output target is made according to the input representing the known 200 song mood. The output target contains 4 rows representing mood and 200 columns for 200 songs. When 5 important features are selected, the goal is to determine how well the multilayer feed to the front of the convolution neural network will classify the songs. The results of the tables we have produced are summarized in Table I below.

TABLE I. CNN RESULT

		Actual			
		Happy	Peaceful	Sad	Angry
Predicted	Happy	43	2	3	2
	Peaceful	2	40	3	5
	Sad	1	2	43	5
	Angry	5	6	1	38
Accuracy		86%	80%	84%	76%

On the table I above shows that the prediction of happy mood is worth 43 of 50 data that has been labelled with a happy mood song. Then on a peaceful prediction yields a value of 40 of 50 data labelled peaceful. Then the prediction result for a sad mood song is 43 out of 50 sadly labelled data. Then for the last test can be seen that the angry song produces a score of 38 from 50 songs labelled angry. From the results we get, we can determine the accuracy of song classification based on the mood using CNN.

V. CONCLUSION

The Convolutional Neural Network classification method is a typical process. Networks must be created with a set of accurate feature values and with an accurate set of network attribute values. The feature extraction is the core of the mood determination of a song that relies heavily on the value feature. A thorough, in-depth and accurate research should be done to identify the best feature sequences that can be extracted from an audio track. Losing accuracy will definitely produce the wrong result.

In this paper, we classified song using CNN. According to research that been conducted, we have the result of accuracy of 82% on average. The highest accuracy in the prediction of happy type music, which is 86% and the lowest is angry type music, that is 76%. In future research can be some upgrade the accuracy by means of some other preprocessing procedures and creating more balanced dataset. Other than that, subsequent research is expected to increase the number of dataset used so that the results obtained will be much more perfect.

REFERENCES

[1] Michael Miller MD, Positive Emotions and the Endothelium: Does Joyful Music Improve Vascular Health University Maryland School of Medicine, 2008.

[2] Yuna Ferguson, Trying to be Happy Can Work: Two Experimental Studies. Knox Collage, 2012.

[3] R. E. Thayer, The Biopsychology of Mood and Arousal. New York Oxford University Press, 1989.

[4] Stuart, GW & Laraia, Principles and Practice of Psychiatric Nursing. St. Louis: Mosby, 2013.

[5] J.A. Russel, A Circumflex Model of Affect. Journal of Personality and Social Psychology, vol. 39, no. 6, pp. 1161 1178, 1980.

[6] Avijeet Kartikay, Harish Ganesan, Vanadana M Ladwani, ”Classification of Music into Moods using Musical Feature. Inventive Computation Technology (ICICT), 2016.

[7] http://cvml.unige.ch/databases/emoMusic/.

[8] Oliver Lartilot, Petri Toivianen, A MatLab Toolbox for Musical Feature Extraction from Audio. Internax, 2007.

[9] Aathreya S. Bhat, Amith V. S, Namrata S. Prasad, Murali Mohan D, An Efficient Classification Algorithm For Music Mood Detection In Western and Hindi Music Using Audio Feature Extraction. International Conference on Signals and Image Processing, 2014.

[10] Krizhevsky A, Sutskever I, Hinton G, ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 201.

[11] Soleymani, Mohammad and Caro, Micheal N. and Schmidt, Erik M. and Sha, Cheng-Ya and Yang, Yi-Hsuan, 1000 Songs for Emotional Analysis of Music. Proc. of the 2Nd ACM International Workshop on Crowdsourcing of Multimedia, 2013.

[12] Eyben, Florian and Weninger, Felix and Gross, Florian and Schuller, Bjorn, Recent Development in openSMILE, the Munich Open-source Multimedia Feature Extractor. ACM International Workshop on Crowdsourcing of Multimedia, 2013.

Comparison Of sGA And SEGA Methods To Solve The Problem Of Power Generation And Power Losses On Distributed Generating Systems

Kukuh Widarsono
Department of Electrical Engineering
Politeknik Negeri Madura
Madura, Indonesia
E-mail: kukuh@poltera.ac.id,kukuh.widarsono@ieee.org

Farid Dwi Murdianto
Department of Electrical Engineering
Electronics Engineering Polytechnic Institute of Surabaya
Surabaya, Indonesia
E-mail: farid@pens.ac.id,fariddm@ieee.org

Akhmad Arif Kurdianto
Department of Electrical Engineering
Politeknik Negeri Madura
Madura, Indonesia
E-mail: aak.smart.technic@gmail.com

M. Sohibul Hajjah
Department of Electrical Engineering
Politeknik Negeri Madura
Madura, Indonesia
E-mail: hajahsohibul@yahoo.co.id

***Abstract*— This paper introduces a new algorithm to solve the problem of power flow optimization. This algorithm is the Spontaneous Evolutionary Genetic Algorithm (SEGA). SEGA is a combination of Neural Network (NN) and Standard Genetic Algorithm (sGA). SEGA conducts individual selection in a population to get the best results. This individual is a representation of each power station in an electrical network system, because the generator is the object to be optimized. SEGA is different from the previous generation of SGA Standard Genetic Algorithm. SEGA generates new populations (excluding the main population), to get more optimal results. Crossover, mutation and recombination are the selection algorithms used in SEGA, to obtain individual or more optimal results. In this simulation is used IEEE 57 bus plant, as well as compared the quality of the solution between the proposed algorithms with sGA. With this algorithm proven that, SEGA able to improve result (Fitness) from sGA. By using SEGA, the cost of generation on an IEEE 57 bus system is cheaper at 110 $ / hr compared to using sGA.**

Keywords— Power Flow; Standard Genetic Algorithm (sGA); Neural Network (NN); Spontaneous Evolutionary Genetic Algorithm (SEGA)

I. INTRODUCTION

Power system is a system that aims to generate, distribute and utilize electrical energy. Electrical energy is generated by the power system through the energy conversion process. The resulting energy is channeled through the transmission network. So that electric power system we can define as the components of electric power that form an integrated system, relation with power generation process. There are three important components that make up the electric power system that is the generation components, transmission components and distribution components.

In generating electrical energy, many limitations must be considered. These limits are the limits of electricity generation capacity, power distribution, and cost. These limits encourage us to create an effective and efficient service system, to meet the needs of electrical energy. Research on power system operation has been done a lot before [1-3], and there is even a paper on the evaluation of the reliability of power systems [4] which provides extensive library surveys for the evaluation of the reliability of power systems.

Various approaches have been made to overcome the power flow optimization problem. Stability-constrained on the optimization of power flow with numerical time-domain based has been proposed [5]. A similar approach is that developing a time-domain dispatch algorithm by including contingencies [6-7] has also been proposed. Further studies taking into account oscillations and stresses in unstable conditions are based on the minimum dynamic stability indices of the singular value "MSV" [8].

Artificial intelligence is widely used to solve electrical problems. artificial intelligence which is often used in solving electrical problems such as Neural-Network (NN) [9-16] and Standart Genetic Algorithm (sGA) [17-18].

This research proposed a combination of Neural Network and Genetic Algorithm standard into a new method of Spontaneous Evolutionary Genetic Algorithm (SEGA) to obtain optimum power flow, with optimum power flow expected to be obtained minimum cost of generation. SEGA assists the computation process in order to obtain the optimum value on each of the power flow parameters.

II. RESEARCH ALGORITHMOLOGY

A. Bus Parameters

In this simulation will be used data channel system IEEE

57 bus. The IEEE 57 bus system consists of 7 generators and 78 channels. The generator is mounted on buses 1, 2, 3, 6, 8, 9 and bus 12. In this system, the generator on bus 1 acts as a generator or slack bus.

The bus types on the IEEE 57 bus system are given as follows.

• One slack bus, means a generating bus on bus 1.
• Six bus generators.
• Fifty bus loads.

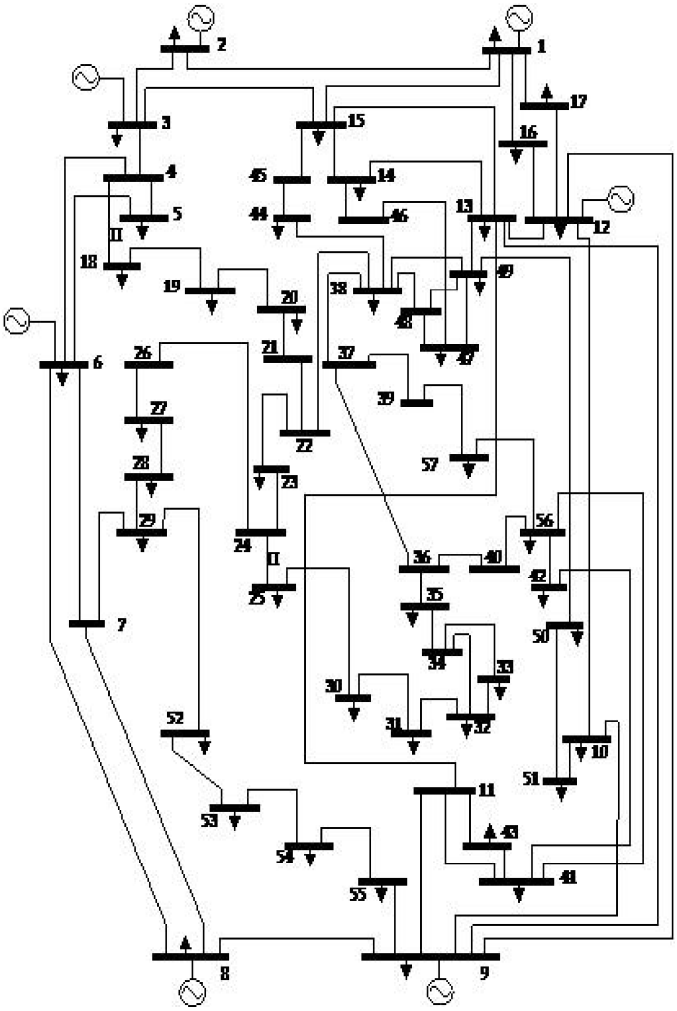


Fig. 1. Single line of IEEE 57 bus

B. Channel Parameters

The other specified bus parameter is Z base. The value of Z base on the channel can be determined using the following formula.

$$Z_{base} = \frac{(\mathrm{kV_{base}})^2}{\mathrm{MVA_{base}}} \quad (1)$$

The value of channel impedance or Z on the system is expressed in units of Ω (Ohm). In order to facilitate the calculation, then the unit must be converted into unit p.u (per unit). Calculation of Z value in p.u unit can be determined using formula

$$Z_{p.u} = \frac{\mathrm{Z_{ohm}}}{\mathrm{Z_{base}}} \quad (2)$$

C. System Modeling Evaluation

After determining the bus and channel parameters, the system is then evaluated by flow power analysis to determine the stability of the system model. For large-scale power systems, the Newton-Rapshon algorithm is very efficient and practical to obtain large currents on the transmission and power in and out of the bus. The complex power form in bus i is expressed by the following formula. [13]

$$\mathrm{P_i} - \mathrm{jQ_i} = |V_i \left|\angle_{\mathrm{i}}^{\delta} \sum_{j=1}^{n} |Y_{ij}||V_j \right|\angle\theta_{ij} + \delta_i \quad (3)$$

Separation on real and imaginary parts then

$$\mathrm{P_i} = \sum_{j=1}^{n} |V_i \| V_j \left|Y_{ij}\right| \cos(\theta_{ij} - \delta_i + \delta_i) \quad (4)$$

$$\mathrm{Q_i} = -\sum_{j=1}^{n} |V_i||V_j|Y_{ij} \left| \sin(\theta_{ij} - \delta_i + \delta_i) \right. \quad (5)$$

$$\begin{bmatrix} \Delta P \\ \Delta Q \end{bmatrix} = \begin{bmatrix} J_1 & J_2 \\ J_3 & J_4 \end{bmatrix} \begin{bmatrix} \Delta\delta \\ \Delta|V| \end{bmatrix} \quad (6)$$

The element J1 is expressed by

$$\frac{\partial \mathrm{P_i}}{\partial \delta_{\mathrm{i}}} = \sum_{j \neq i} |V_i \| \mathrm{V_j}|Y_{ij}| \sin(\theta_{ij} - \delta_i + \delta_i) \quad (7)$$

$$\frac{\partial \mathrm{P_i}}{\partial \delta_{\mathrm{i}}} = -|V_i||V_j|Y_{ij}| \sin(\theta_{ij} - \delta_i + \delta_i) \qquad \mathrm{j}\neq\mathrm{i} \quad (8)$$

The element J2 is expressed by

$$\frac{\partial \mathrm{P_i}}{\partial |\mathrm{V_i}|} = 2|V_i||Y_{ii}|cos\theta_{ii} + \sum_{j \neq i} |V_j \| \mathrm{Y}_{ij}| \cos(\theta_{ij} - \delta_i + \delta_i) \quad (9)$$

$$\frac{\partial \mathrm{P_i}}{\partial |\mathrm{V_j}|} = |V_i||Y_{ij}| \cos(\theta_{ij} - \delta_i + \delta_i) \qquad \mathrm{j}\neq\mathrm{i} \quad (10)$$

Element J3 is expressed by

$$\frac{\partial \mathrm{Q_i}}{\partial \delta_{\mathrm{i}}} = \sum_{j \neq i} |V_i \| \mathrm{V_j}|Y_{ij}| \cos(\theta_{ij} - \delta_i + \delta_i) \quad (11)$$

$$\frac{\partial \mathrm{P_i}}{\partial \delta_{\mathrm{i}}} = -|V_i \| \mathrm{Y}_{ij}| \cos(\theta_{ij} - \delta_i + \delta_i) \qquad \mathrm{j}\neq\mathrm{i} \quad (12)$$

Element J4 is declared with

$$\frac{\frac{\partial \mathrm{Q_i}}{\partial |\mathrm{V_i}|}}{\frac{\partial \mathrm{Q_i}}{\partial |\mathrm{V_i}|}} = -2|V_i \| \mathrm{Y}_{ii}| \sin\theta_{ii} - \sum_{j \neq i} |V_j \| \mathrm{Y}_{ij}| \sin(\theta_{ij} - \delta_i + \delta_i) \quad (13)$$

$$\frac{\partial \mathrm{Q_i}}{\partial |\mathrm{V_j}|} = -|V_i||Y_{ij}| \sin(\theta_{ij} - \delta_i + \delta_i) \qquad \mathrm{j}\neq\mathrm{i} \quad (14)$$

$\Delta P_i^{(k)}$ and $\Delta Q_i^{(k)}$ is the difference between the value entered and the calculation value called residual power

$$\Delta P_i^{(k)} = P_i^{sch} - P_i^{(k)} \quad (15)$$

$$\Delta O_i^{(k)} = O_i^{sch} - O_i^{(k)} \quad (16)$$

The new estimate for bus voltages is

$$\delta_i^{(k+1)} = \delta_i^{(k)} - \Delta\delta_i^{(k)} \quad (17)$$

$$\left|V_i^{(k+1)}\right| = \left|V_i^{(k)}\right| + \Delta\left|V_i^{(k)}\right| \quad (18)$$

The iteration process will stop if the following equation has been met

$$\left|\Delta P_i^{(k)}\right| \leq \varepsilon \quad (19)$$

$$\left|\Delta Q_i^{(k)}\right| \leq \varepsilon \quad (20)$$

Optimization Using Lagrange Algorithm

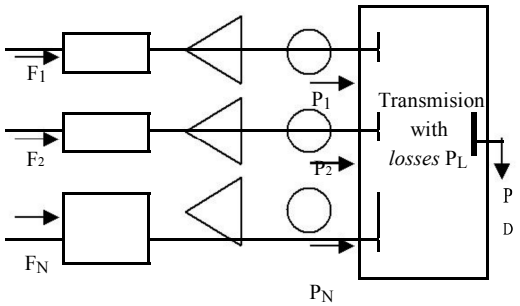


Fig. 2. N Thermal Generating Unit Serving Load PD

Mathematically, the problem can be written in theobjective function stated in the FT which is equal to thetotal costsupplying load and loss in transmission.

$$F_T = F_1 + F_2 + F_3 + \cdots + F_N = \sum_{i=1}^{N} \mathrm{F(P_i)} \quad (21)$$

Solving the problem is how to minimize FT withtaking into account the limitation that the amount of power generated must be equal to the load power including the transmission loss. The mathematical form is

$$\min \sum_{i=1}^{N} \mathrm{F(P_i)} = \min \sum_{i=1}^{N} (\alpha_i + \beta_i P_i + \gamma_i P_{i^2}) \quad (22)$$

Lagrange equation can be written

$$L = F_T + \lambda\varphi \quad (23)$$

$$\frac{\partial L}{\partial P_i} = \frac{\partial F_i}{\partial P_i} + \lambda\left(\frac{\partial P_D}{\partial P_i} + \frac{\partial P_L}{\partial P_i} - \frac{\partial P_I}{\partial P_i}\right) = 0 \quad (24)$$

$$\frac{\partial L}{\partial P_i} = \frac{\partial F_i}{\partial P_i} + \lambda\left(\frac{\partial P_L}{\partial P_i} - 1\right) = 0 \quad (25)$$

$$\frac{\partial L}{\partial P_i} = \frac{\partial F_i}{\partial P_i} - \lambda\left(1 - \frac{\partial P_L}{\partial P_i}\right) = 0 \quad (26)$$

$$\frac{\partial F_i}{\partial P_i} = \lambda\left(1 - \frac{\partial P_L}{\partial P_i}\right) \quad (27)$$

With the following information:

- P_D it doesn't depend on change by P_i, so

$$\frac{\partial P_D}{\partial P_i} = 0$$

- P_L changed depending on value of current and impedance network, so

$$\frac{\partial P_L}{\partial P_i} \neq \frac{dP_L}{dP_i}$$

Figure 3 shows the flow chart of the algorithm Spontaneous Evolutionary Genetic Algorithm (SEGA). The steps to get an optimum power flow on the system are:

1. Generating population at random, where the width of the population is adjusted by the number of objects to be in optimization. In this case the object in the optimization is the generator.

2. Chromosome encoding is the conversion of the assumption value to the real value. This value is represented by the randomly generated chromosome. While the real value is the value of the real power generation.

3. Crossover, mutation and recombination algorithms are used to select the most optimal individuals in the population.

4. The process will continue to recur to get the best chromosome. The best chromosomes are represented in the form of fitness values.

5. In certain generations, if Fitness selection is less than standard. So SEGA will generate a new population, to replace the old population.

6. The new population is obtained from the results of Neural Network training on the best individual sGA results. With this expected to emerge a homogeneous population of quality.

7. The best chromosome of the whole process will appear after the maximum iteration has been met.

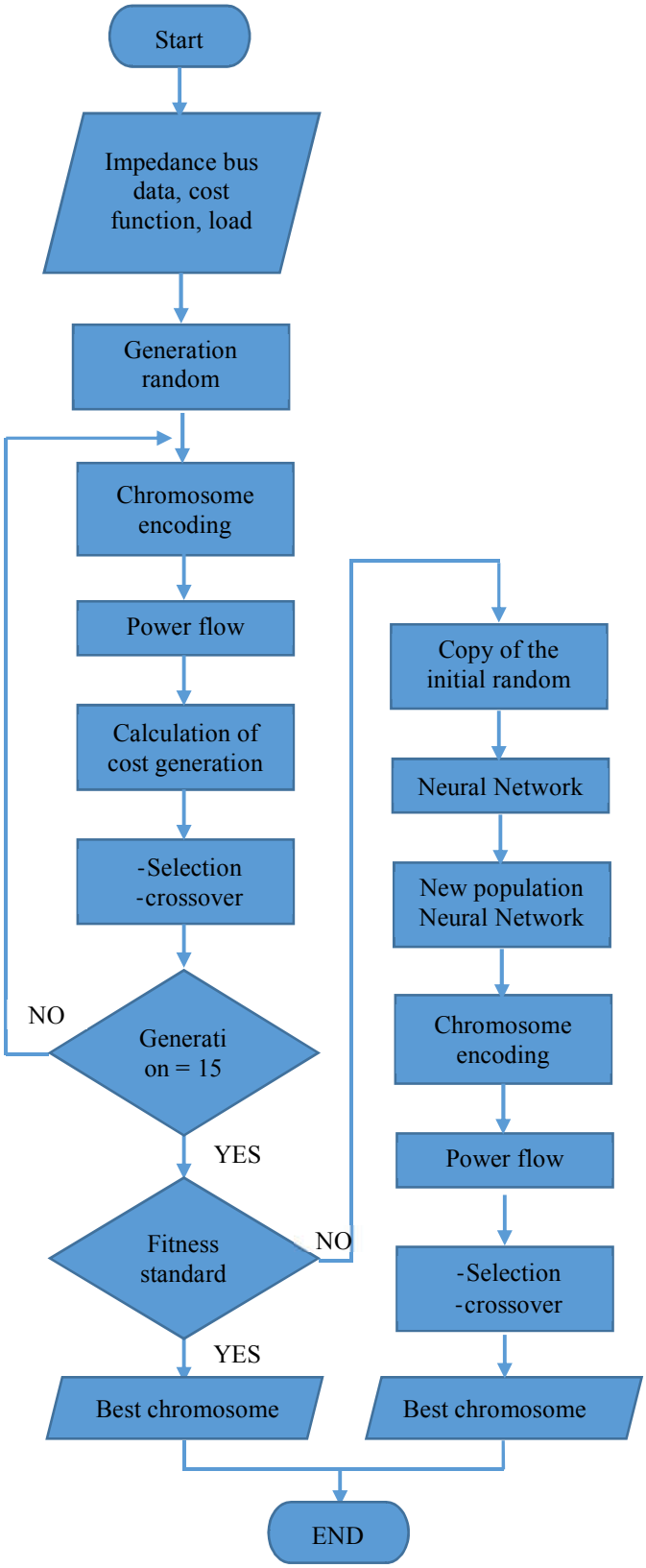


Fig. 3. Flowchart of SEGA Algoritm

III. SIMULATION RESULTS

Comparison of the amount of power generated on the IEEE 57 bus system between the sGA and SEGA algorithms are shown in Table I.

TABLE I. POWER GENERATION AND LOSSES

The amount of Power Generation and Losses				
No	sGA		SEGA	
	PG (MW)	PL (MW)	PG (MW)	PL (MW)
1	1.272.966	22.183	1.269.884	19.101
2	1.270.836	20.053	1.270.246	19.463
3	1.271.943	21.16	1.270.227	19.444
4	1.272.252	21.47	1.270.351	19.569
5	1.270.155	19.372	1.270.294	19.511
6	1.271.991	21.209	1.269.914	19.131
7	1.272.078	21.296	1.270.489	19.706
8	1.272.966	22.183	1.270.475	19.692
9	1.270.836	20.053	1.270.292	19.509
10	1.271.943	21.16	1.270.722	19.939

The results of 10 simulations of the IEEE 57 bus system are shown in table 1, the SEGA algorithm is capable of generating lower power than the sGA with a difference of 2 MW. Only on 5 SEGA simulation is bigger than sGA, but the difference is not too big. Comparation Algorithm sGA and SEGA on power generation are shown in Fig.4.

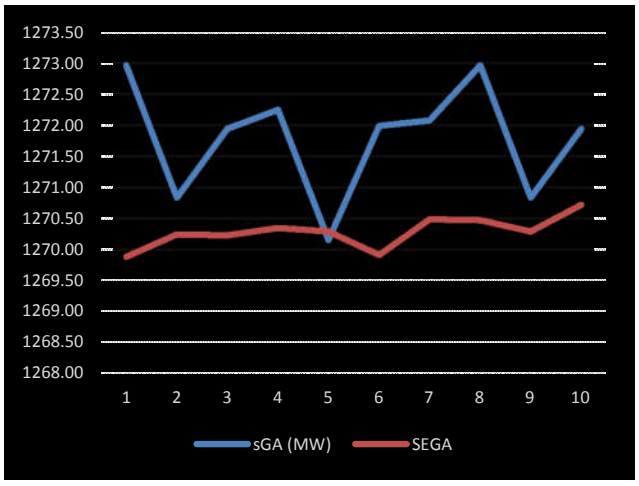


Fig. 4. Comparation of power generation (MW)

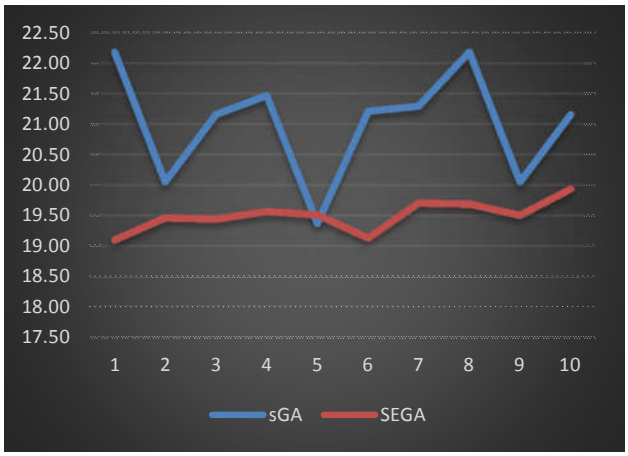


Fig. 5. Comparation of power losses (MW)

From table I there is a difference between power generation and load power, this can be seen from the amount of PL (losses / P loss). These power losses are derived from transmission line losses. Comparation Algorithm sGA and SEGA on power loss are shown in Fig.5.

Comparison of the amount of cost power generation of the IEEE 57 bus system between the sGA and SEGA algorithms are shown in Fig.6.

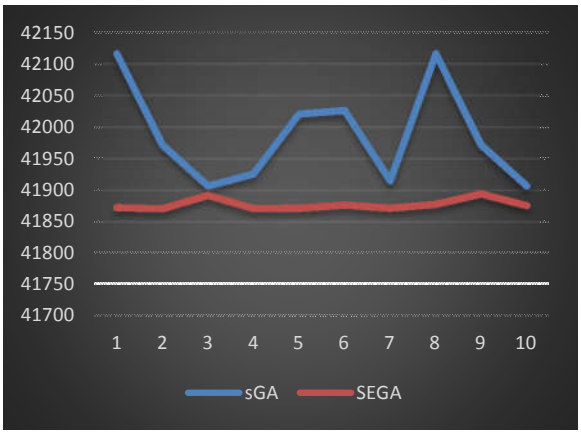


Fig. 6. Comparation of cost generation ($)

The generating cost of the IEEE 57 bus compare the 2 algorithms tested by the author. It is known that SEGA is able to provide a cheaper cost of generation compared to sGA. That is the difference in cost about 110 $ / hr.

Figure 7 shows the simulated fitness result between SEGA and sGA. With the same population SEGA generated fitness is better than sGA. In the 15th generation, Fitness SEGA increased after stagnating at 0.2378 point. This shows that, the new population of NN learning results work in the 15th generation. This population replaces the old population of the previous generation. This population is aimed at improving the fitness of sGA selection results.

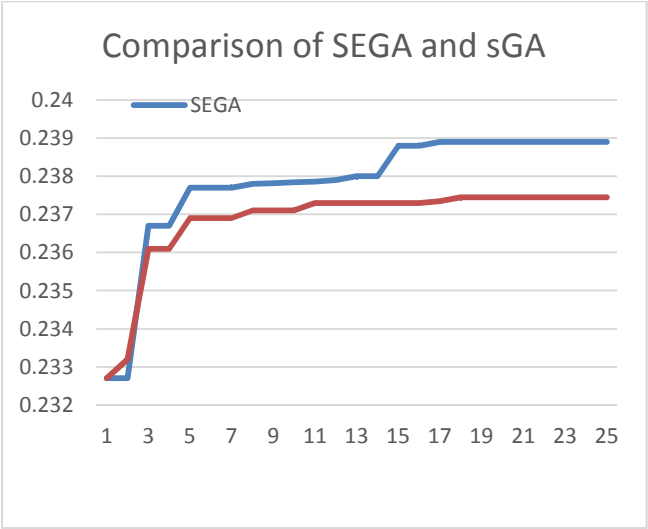


Fig. 7. Fitness running IEEE 57 bus between SEGA with sGA

The SEGA Algorithm result is better than sGA in cost of power generation and power losses are shown in table 1 and Fig. 6-7. The best results are obtained from the new population generated by this SEGA algorithm. The individual population has a quality chromosome. These chromosomes are the result of training from the neural network, where the target of this training is the best individual outcomes of the sGA (after 30 sGA simulations and the best one). While the individual population of sGA is generated randomly, so the result is a heterogeneous population. With a heterogeneous population, the resulting chromosome is also not necessarily qualified.

IV. Conclusion

From the simulation results can be concluded that the SEGA algorithm capable of generating lower generation power and lower cost compared to sGA. With a lower 2 MW power and a cost of 110 $ / hr is cheaper than sGA.

REFERENCES

[1] Fumihiro Endo, Masaki Kanamitsu, Ryo Shiomi, Hiroki Kojima, Naoki Hayakawa, and Hitoshi Okubo, “Optimization of Asset Management and Power Sistem Operation Based on Equipment Performance”, Interenational Conference on Condition Monitoring and Diagnosis, Beijing, China, April 21-24, 2008.

[2] A.A. Abou El-Ela, T. Fetouh, M.A. Bishr, and R.A.F. Saleh, “Power sistem Operation using Particle Swarm Optimization Technique”, Science Direct - Electric Power Sistem Research, Vol.78, 2008, pp.1906-1913.

[3] A. Shunmugalatha and S. Mary Raja Slochanal, “Optimum Cost of Generation for Maximum Loadability Limit of Power Sistem using Hybrid Particle Swarm Optimization”,Science Direct – Electrical Power and Energy Sistems, Vol.30, 2008, pp.486-490.

[4] R. Bilinton, Mahmud Fotuhi, Firuzabad, and Lina Bertling, “Bibliography on the Application of Probability Methods in Power Sistem Reliability Evaluation 1996-1999”, IEEE Transaction on Power Sistems, Vol.16, No. 4, November 2001

[5] D. Gan, R. J. Thomas, dan R. D. Zimmerman, “Stability-constrained optimal power flow”, IEEE Trans. Power Syst., vol. 15, no. 2, pp. 535–540, May 2000.

[6] S. Bruno, E. D. Tuglie, dan M. La Scala, "Transient security dispatch for the concurrent optimization of plural postulated contingencies", IEEE Trans. Power Syst., vol. 17, no. 3, pp. 707–714, Aug. 2002.

[7] S. K. M. Kodsi dan C. A. Cañizares, "Application of a stability constrained optimal power flow to tuning of oscillation controls in competitive electricity markets," IEEE Trans. Power Syst., vol. 22, no. 4, pp. 1944–1954, Nov. 2007.

[8] B. Jayasekara dan U. Annakkage, "Derivation of an accurate polynomial representation of the transient stability boundary," IEEE Trans. Power Syst., vol. 21, no. 4, pp. 1856–1863, Nov. 2006.

[9] Gui M., Pahwa A., Das S., "Analysis of Animal-Related Outages in Overhead Distribution Systems With Wavelet Decomposition and Immune Systems-Based Neural Networks," IEEE Transactions on Power Systems, Vol. 24, no. 4, November 2009.

[10] Kazemlou S., Mehraeen S., "Decentralized Discrete-Time Adaptive Neural Network Control of Interconnected DC Distribution System Derivation of an accurate polynomial representation of the transient stability boundary," IEEE Transactions on Smart Grid, Vol. 5, no. 5, September 2014.

[11] Guo M.F, Zeng X.D, Chen D.Y, Yang N.C, "Deep-Learning-Based Earth Fault Detection Using Continuous Wavelet Transform and Convolutional Neural Network in Resonant Grounding Distribution Systems," IEEE Trans. Power Syst., vol. 21, no. 4, pp. 1856–1863, Nov. 2006. IEEE Sensors Journal, Vol. 18, no. 3, February 1, 2018.

[12] Manitsas E., Singh R., Pal B.C, Strbac G., "Distribution System State Estimation Using an Arti fi cial Neural Network Approach for Pseudo Measurement Modeling," IEEE Transactions on Power Systems, Vol. 27, no. 4, November 2012.

[13] Ding N., Benoit C., Foggia G., Besanger Y., Wurtz F., "Neural Network-Based Model Design for Short-Term Load Forecast in Distribution Systems," IEEE Transactions on Power Systems, Vol. 31, no. 1, January 2016.

[14] Nakhli M.P., Savafi A.A, "Path Characteristic Frequency-Based Fault Locating in Radial Distribution Systems Using Wavelets and Neural," IEEE Transactions on Power Delivery, Vol. 26, no. 2, April 2011.

[15] Li H., Misra S., "Prediction of Subsurface NMR T2 Distributions in a Shale Petroleum System Using Variational Autoencoder-Based Neural Networks," IEEE Geoscience and Remote Sensing Letters, Vol. 14, no. 12, December 2017

[16] Heidari A., Agelidis V.G., Pou J., Aghaei J., Ghias A.M.Y.M, "Reliability Worth Analysis of Distribution Systems Using Cascade Correlation Neural," IEEE Transactions on Power Systems, Vol. 33, no. 1, January 2018.

[17] Chang C.I., Chen S.Y., Li H.C., Chen H.M., Wen C.H., "Comparative Study and Analysis Among ATGP, VCA, and SGA for Finding Endmembers in Hyperspectral Imagery," IEEE Journal of Selected Topics in Applied Earth Observations And Remote Sensing, Vol. 9, no. 9, September 2016

[18] Jia Y., Andrieu C., Piechocki R.J., Sandell M., "Depth-First and Breadth-First Search Based Multilevel SGA Algorithms for Near Optimal Symbol Detection in MIMO Systems," IEEE Transactions on Wireless Communications, Vol. 7, no. 3, March 2008.

The using of Object-Oriented Method to Developing Android Application for Mapping Distribution of MSMEs (Micro, Small Medium Enterprises)

Addin Aditya
Department of Information Systems
STIKI Malang
Jl. Raya Tidar No 100, Malang City 65146, East Java, Indonesia
addin@stiki.ac.id

***Abstract*—The objective of this research is to develop a Mobile Geographic Information System based on android mobile from user and customer requirements, which it will help the business owner in marketing their products. MSMEs has a big contribution not only to economic growth but also become a private income source for most of the people in Indonesia. Therefore, to increase the independence, efficiency and effectivity of MSMEs, utilization of information technology should be improved. The reason why a mobile application was chosen because mobile technology is one of technology that often used by the community. So, it's easy to use. An object-oriented approach was chosen in this system development because it can give a clear and understandable information to the stakeholders and also it can describe the task and function of each actor in the system. We expect that this model will help the developer to see the system and application comprehensively. And also help the owner of MSMEs to increase their effectiveness and efficiency in promoting their product**

Keywords—android application, MSMEs, location, object-oriented modeling

I. INTRODUCTION

In order to reach national development that to realize a prosperous society as stated in Pancasila, Indonesia's Government is always maximized the utilization of the natural or human resource. To raise the economical of Indonesia, the government give several programs such as society empowerment and also small business to poverty countermeasures. In this case, Micro, Small and Medium Enterprises (MSMEs) has a good potential because it has a big contribution in absorbing more than 99.45% manpower and 30% contribution for Gross Domestic Product (GDP) in Indonesia [1]. After the crisis in 1997/1998, the government starting to give attention to MSMEs development either in quantity or quality. These things happened because MSMEs has a good sustainability instead of large business. These phenomena explain that MSMEs is a productive sector to developt for supporting the development of Indonesia's economic either macro or micro.

Table 1. The Contribution of MSMEs and Large Business

Year	2008	2009	2010	2011	2012
Unit					
MSMEs	51409612	52764603	53283732	55206444	56534592
Large Business	4650	4677	4838	4952	4968
Manpower					
MSMEs	94024278	96211332	99401775	1010772458	106657509
Large Business	2756205	2674671	2839711	2891224	3150645
GDP (Million Rupias)					
MSMEs	2613226	2993151	3466393	4303571	4869568
Large Business	2080582	2301709	2602369	3123514	3372296

One of the obstacles that faced by MSMEs owner is lack of capital, lack of managerial capability and limitation of the market. Besides, there are unhealthy competition and also an economic insistence that caused limitation of scope [2]. This concern was based on fact that Indonesia will face MEA and a free market. When it happened, MSMEs should have to keep maintain their business.

To expand market share and increasing competitiveness of MSMEs, they need an application that can integrate and automate the business process of MSMEs. Several critical issues that faced by MSMEs is lack of capital and information technology capabilities. In order to understand the requirement of MSMEs services and keep it in the right direction, we need a model that can describe the need of MSMEs and an application that can connect it with the society. In this paper, we using an object-oriented as a method for modeling the business processes of MSMEs. Object-oriented system development is a way to develop software by building self-contained modules or objects that can be easily replaced, modified and reused. It depicts the view of the real world as a system of cooperative and collaborative objects [3].

II. LITERATURE REVIEW

A. Micro, Small Medium Enterprises (MSMEs)

Several facts founded in some many research. MSMEs could become preventive solutions to poverty in Indonesia by developing a potential small economic sector. A program to poverty alleviation through MSMEs empowerment has been planned by President of Indonesia on February 26 2005. There is four main activity in it, they are:

1. Growing up a conducive business climate
2. Developing of a business support system
3. Developing of entrepreneur
4. Empowerment of microscale business

In 2008, MSMEs in Indonesia has given 55.56% contribution for GDP based on Ministry of Indonesia Cooperation and Small Medium Enterprises in [4]. MSMEs as part of economics should be more innovative. Competitive advantage based on innovation and creativity should be declared as a priority because it has a sustainable durability and long time period [5]. The Problem in productivity of MSMEs is lack concern of rights of creativity and innovation. This problem causing product copying so it will break the creator of the product. Property right of product or design product which is abandoned will be creating product incentive. Therefore, we need an appreciation by giving a patent for innovative MSMEs. This will be encouraging another creation and also produce a product with a better design and result to an interesting customer.

According to Indonesia Constitution No 20 clause 1, 2008 about Micro, Small and Medium Business, there are several definitions of MSMEs:

1. A Micro business is a productive business that owned by an individual or organization that fulfill the criteria as arranged in constitutions
2. Small business is a private productive economic business that owned by owned by an individual or organization and not part of other company
3. Medium business is standalone productive business, by personal or organizations that not apart of another large business
4. A Large business is a productive economic business that owned by organizations with an annual net worth more than medium business
5. The Business world is micro, small and large businesses that do economic activity in Indonesia and stay in it.

Table 2. Business Criteria based on Assets and Turnover

Business Scale	Criteria	
	Asset	Turnover
Micro Business	Rp 50 million max	Rp 300 million max
Small Business	More than Rp 50 million or equals to Rp 500 million	More than Rp 300 million or equals to Rp 2.5 billion
Medium Business	More than Rp 500 million or equals to Rp 10 billion	More than Rp 2.5 billion or equals to Rp 50 billion
Large Business	More than Rp 10 billion	More than Rp 50 billion

Moreover, MSMEs has their own characteristics based on their product commodity [6], there are:

1. Because most of MSMEs couldn't have a proper technology, they produce handmade products so that they have a various quality of a product
2. They have a limitation of the design product. Most of MSMEs work by order and not interested in taking a risk to make a new design
3. Lack of variety product. Usually, MSMEs just produce few kinds of product
4. Limitation of price list and product capacity
5. unstandardized of product materials
6. unguaranteed of product continuity

B. MSMEs Strategy and Business Opportinity

Development can be defined as an effort to raise conceptual ability, theory, technical and individual morals. The impact of MSMEs development in Indonesia and the government roles in order to raise the growth of MSMEs has a positive result, either directly or indirectly [7]. Basically, the development of MSMEs has shared responsibility between the government and the society. By observing many problems that faced by MSMEs owners, several efforts should be completed as explained in Indonesia Constitutions.

The important role of MSMEs is not only for economic growth in big cities but also in the rural area. Here is several important roles of MSMEs:

1. MSMEs has an important role to provide an economic service widely to the society
2. MSMEs help the government to reduce unemployment by creating new jobs
3. MSMEs more flexible than large business
4. MSMEs does not depend on the import component and always use a local resource, therefore it can save foreign exchange.
5. A Place for new entrepreneurs
6. MSMEs is an instrument of income equity and decrease the equity of public welfare

Table 3. Strategic Aspect of MSMEs

Politic Aspect	Economical Aspect	Social Aspect
The economic development between countries especially in ASEAN and APEC will be creating a new opportunity for MSMEs	➢ MSMEs has a contribution to GDP forms ➢ The development of MSMEs is only needed a low level of investment ➢ MSMEs has a contribution in non-oil export	➢ MSMEs sector has been ensured the stability of labor market ➢ Reduction of unemployment will be raising new entrepreneurs

C. Object-Oriented Modeling

Object-oriented analysis and design (OOAD) is the methods that leads us to an object-oriented decomposition. By applying object-oriented design, we can create software that resilient to change and written with an economy of expression [8]. The Object can be defined as an entity that has an identity, state and behavior. There are several common principals of OOAD:

1. Model the context: useful system that corresponding with OOAD context
2. Reuse patterns: built based on powerful ideas and pretested component fixing system quality and productivity from a development process
3. Tailor the method to suit specific projects: every step of the development process has a unique challenges OOAD must fit with requirements from given analysis and design
4. Emphasize the Architecture: understandable architecture that facilitating collaboration between designer and programmer

In the object-oriented approach, the focus is on capturing the structure and behavior of information systems into small modules that combine both data and process [9]. The main aim of the object-oriented design is to improve the quality and productivity of system analysis and design by making it more usable. In the analysis phase, object-oriented models are used to fill a gap between problem and solution. It performs well in situations where the systems are undergoing continuous design, adaption and maintenance. It identifies objects in a problem domain, classifying them in terms of data and behavior.

Using object-oriented analysis, it may be possible improved reliability and flexibility, code reusability, reduce maintenance cost and also real-world modeling [10]. This method does have the capability to minimize few major operating expenses related with the model or structure, such as system may essentially require maintenance and also continue development of the general programming template and code data structures.

III. Proposed Model

A. Research Methodology

The resources of this study are taken from MSMEs and Cooperative Service in Surabaya city period of April 2017. Prototyping model has been chosen in this paper as an application development method where it can be used to connecting client incomprehension about technical aspect and clearing the client requirement analysis to developers.

Prototyping method starts with collecting client requirement for application. Then, the prototype model will be built in order to make client can imagine what they want easily. This method will provide an interface with flow simulation so that it can look like a real software. Finally, the prototype program will be evaluated by a user until user specification is defined.

B. Requirement Analysis

The requirement analysis starts with identifying the actors between MSMEs business actor, Cooperative and MSMEs service and the society. It will show in the table below:

Table 4. Collection of Application Requirement

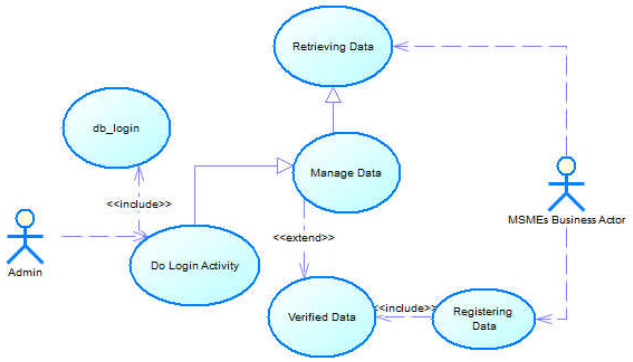
Actor	Requirements	System
MSMEs (Owners)	Promote the product	Using map / location-based services as bookmark location
	Information dissemination of MSMEs	Detail information of their business can get easily
Cooperative and MSMEs service	Increase the GDP	Data visualizations
	Society service	Accessible of MSMEs information
Society (User)	MSMEs Information	Data of MSMEs based on location in 2D map in mobile apps

C. Modeling System

1) Use Case Model

This model has an important role to modeling and organizes the behavior of a system. From figure 1 can be seen how a developer can describe each process of actors and its relationship.

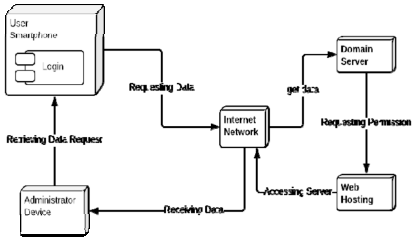
Figure 1. Use-Case Model Application



2) Deployment Diagram

User flow process of the business through smartphone while administrator from a personal computer.

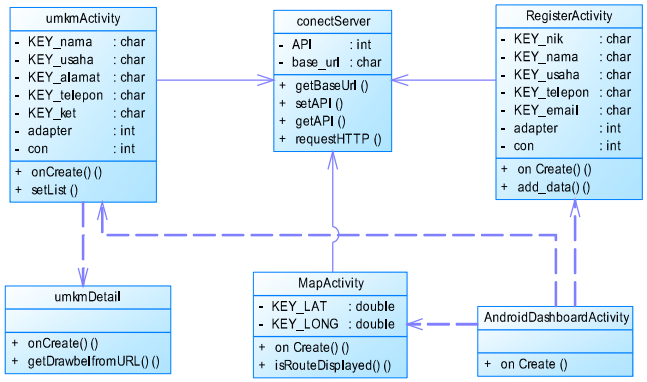
Figure 2. Deployment Diagram Estimation



3) Class Diagram

This diagram provides a graphical view of the classes, interfaces and packages that compose a system and the relationship between them. Here is a class diagram model for MSMEs mapping distribution application.

Figure 3. Class Diagram Model



IV. RESULT AND DISCUSSION

A. Software Testing

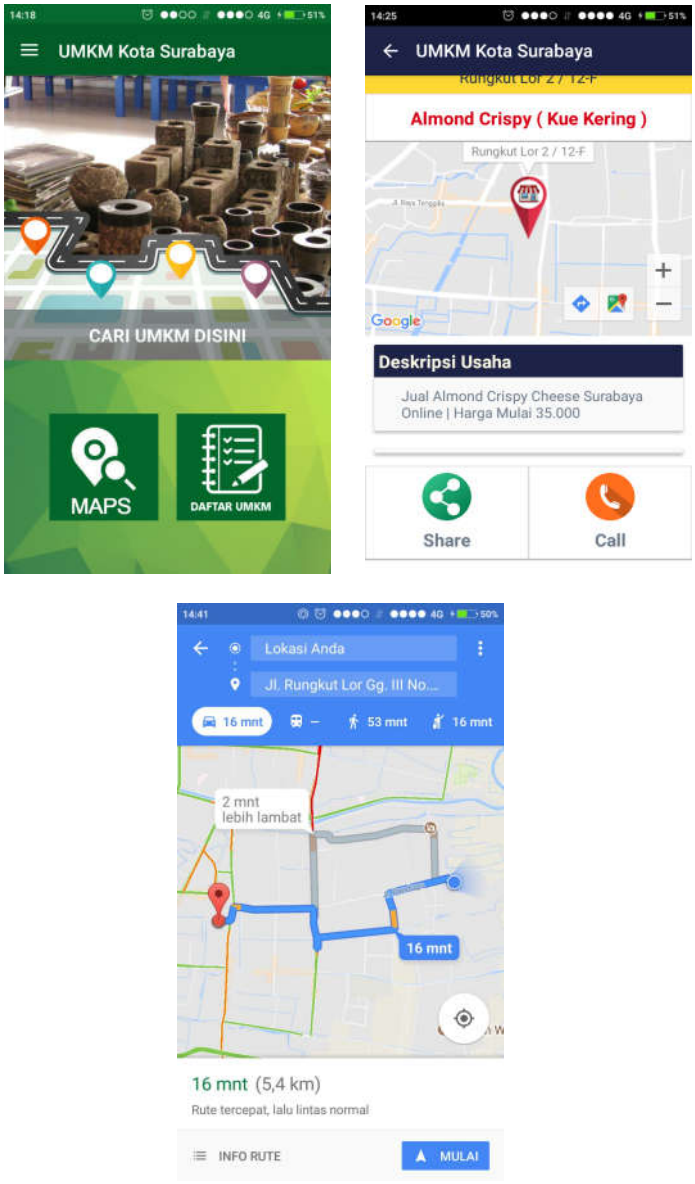
Testing is the process of evaluating a system or its component(s). according to ANSI/IEEE 1059 Standard, testing can be defined as a process of analyzing a software item to detect the difference between existing and required conditions and to evaluate the features of the software items.

In this paper, we using black box testing as a method. Black box focuses on testing the functionality of the system using the functional requirements [11]. Here are functional requirements and expected outcome from an application:

Table 5. Functional Requirements and Expected Outcome

Test ID	Purpose	Input	Expected Outcome	Status
1	Displaying the main menu	Choosing icon	Displaying the main menu of an application	[ ] fit [ ] doesn't fit
2	Displaying the details of MSMEs whether in text or in 2D maps	Choosing icon	Displaying details of MSMEs in text and image	[ ] fit [ ] doesn't fit
			Displaying the location of MSMEs in 2D maps	[ ] fit [ ] doesn't fit
			Displaying the route of MSMEs based on user's location	[ ] fit [ ] doesn't fit
3	Maintenance Data	Choosing Icon	Displaying the data of MSMEs	[ ] fit [ ] doesn't fit
			Register new data of MSMEs	[ ] fit [ ] doesn't fit
			Delete / updating data	[ ] fit [ ] doesn't fit

Figure 4. User Interface of Application



B. Software Evaluation

The Tested system will be evaluated based on software testing. The purpose of the evaluation is to measure how well the system can provide the user requirement. Table 6 shows how the system can promote the product of MSME. Based on the result can be concluded that users totally agree that the system is very helpful for promoting the product. Table 7 and 8 shows that the system helps for social services and information spreading. And the rest shows that the apps could help increase the GDP and fulfill the information needs.

Table 6. Evaluation of Product Promotion

Answer Code	**Totally agree (TA)**	**Agree (A)**	**Disagree (D)**
Frequent Answer	14	30	6
Score	14 x 3 = 42	30 x 2 = 60	6 x 1 = 6
Total	108		
Percentage	108 / 150 x 100% = 72 %		

Table 7. Evaluation of Information Spread

Answer Code	**SS**	**S**	**TS**
Frequent Answer	24	23	3
Score	24 x 3 = 72	23 x 2 = 46	3 x 1 = 3
Total	121		
Percentage	121 / 150 x 100% = 80 %		

Table 8. Society Services

Answer Code	**TA**	**A**	**D**
Frequency	12	29	9
Score	12 x 3 = 36	29 x 2 = 58	9 x 1 = 9
Total	103		
Percentage	103 / 150 x 100% = 68 %		

Table 9. Increasing GDP

Answer Code	**TA**	**A**	**D**
Frequency	4	27	19
Score	4 x 3 = 12	27 x 2 = 54	19 x 1 = 19
Total	85		
Percentage	85 / 150 x 100% = 56 %		

Table 10. Required Information

Answer Code	**SS**	**S**	**TS**
Frequency	20	25	5
Score	20 x 3 = 60	25 x 2 = 50	5 x 1 = 5
Total	115		
Percentage	115 / 150 x 100% = 76 %		

V. CONCLUSION

According to the evaluation, it can be concluded that the application helps the business owner specially to promoting their product and spread its information. Furthermore, if the system can promote their product by raising the people's purchasing, it will raise the Gross Domestic Product, considering MSME has a big contribution to it. Most the people find it helpful with the system. With this apps, the government can use it for mapping the distribution of MSME. We hope to establish good cooperation between the government and the business owner. However, there are still many things that need to be improved in the apps. Especially in improving communication and collaboration between society, government and business owner.

REFERENCES

[1] Supriyanto, "PEMBERDAYAAN USAHA MIKRO, KECIL, DAN MENENGAH (UMKM) SEBAGAI SALAH SATU UPAYA PENANGGULANGAN KEMISKINAN," *Jurnal Ekonomi dan Pendidikan,* vol. 3, no. 1, pp. 1-16, 2006.

[2] Y. R. Suci, "Perkembangan UMKM (Usaha Mikro Kecil Menengah) di Indonesia," *Jurnal Ilmiah Cano Economics,* vol. 6, no. 1, pp. 51-58, 2017.

[3] R. Nath and D. Kumar, "Introduction to Object-Oriented Methodology," in *Introduction to Object-Oriented Analysis and Design*, 2018, pp. 1-23.

[4] J. W. Saputro, P. W. Handayani, A. N. Hidayanto and I. Budi, "Peta Rencana (ROADMAP) Riset Enterprise Resource Planning (ERP) dengan Fokus Riset pada Usaha Kecil Menengah (UKM) di Indonesia," *Journal of Information Systems ,* vol. 6, no. 2, pp. 140-145, 2010.

[5] P. D. S. U. B. I. d. K. (. P. P. R. T. I. D. Kreativitas), "Darwanto," *Jurnal Bisnis dan Ekonomi,* vol. 20, no. 2, pp. 142-149, 2013.

[6] L. P. P. Indonesia, Profil Bisnis Usaha Mikro, Kecil dan Menengah, Jakarta: Lembaga Pengembangan dan Perbankan Indonesia, 2015.

[7] T. Tambunan, UMKM di Indonesia, Bogor: Ghalia, 2009.

[8] G. Booch, Object-Oriented Analysis and Design, Santa Clara, California: Addison-Wesley, 1994.

[9] Mohtasim, "Object Oriented Approach," Tutorials Points, 2018. [Online]. Available: https://www.tutorialspoint.com//system_analysis_and_design/system_analysis_and_design_object_oriented_approach.htm. [Accessed 18 April 2018].

[10] M. Mukherjee, "Object-Oriented Analysis and Design," *International Journal of Advance Engineering and Management,* vol. 1, no. 1, pp. 18-24, 2016.

[11] H. Freeman, "Software Testing," in *IEEE Instrumentation & Measurement Magazine*, 2002, pp. 48-50.

RESTful Web Service Optimization with Compression and Encryption Algorithm

Achmad Solichin
Faculty of Information Technology
Universitas Budi Luhur
Jakarta, Indonesia
achmad.solichin@budiluhur.ac.id

Mochammad Andika Putra
Informatics Department
Universitas Budi Luhur
Jakarta, Indonesia
mochammadandika9@gmail.com

Khaeri Diniari
Informatics Department
Universitas Budi Luhur
Jakarta, Indonesia
dinikhaeridiniari@gmail.com

***Abstract*—Along with the increasing number of mobile-based applications, the use of web services as a method of data exchange between systems also increased. With the web service technology, the exchange of data between systems with each other becomes easier even though it has a different platform. Increasing the amount of data requires more storage space on the server. Therefore, this research proposed the compression and encryption algorithm to optimize RESTful web service. We proposed LZW compression to reduce the storage space needed to store data on the server. We also proposed a Blowfish and AES-128 algorithm for securing data exchange through the internet network. The results show that all services made can run well and can be accessed through the Android application. In addition, the implementation of compression algorithm resulted in storage space savings of 38% with a compression rate of 1.78. The application of compression algorithm has an impact on the increase of computation time, although not significant.**

Keywords—RESTful Web service, optimization, LZW compression, Blowfish encryption, AES-128 encryption

I. INTRODUCTION

Mobile device users in the world are increasing from time to time. Statista predicts the number of mobile device users in 2018 as much as 2.53 billion people or 36% of the total world population [1]. The StatCounter site also notes that in March 2018 51.43% of internet users accessed the site via mobile devices [2]. The growing number of mobile device users is driving the growth of the number of mobile-based applications and web-based data-service exchange services. With the web service technology, the exchange of data between systems with each other becomes easier even though it has a different platform.

Currently, the standard web-based data exchange service boils down to two standards: SOAP and REST [3]. SOAP stands for Simple Object Access Protocol. SOAP is a simple XML-based protocol to let applications exchange data or information over HTTP. The standard protocol HTTP makes it easier for the SOAP model to pass through firewalls and proxies without any modifications to the SOAP protocol. However, the lack of SOAP is the slowness of data transfer. Some recent research still uses SOAP-based web services but is optimized with compression algorithms to speed up data exchange [4]–[6]. They use compression algorithms like Tagged Sub-optimal Code (TSC), Huffman and J-bit encoding.

Meanwhile, REST stands for Representational State Transfer. REST is a set of principles for data exchange architecture through standard HTTP protocol. REST does not contain any additional messaging layers and focuses on design rules to create services without territory constraints. The client can access the resources using a unique URI, and the resource representation is returned. When accessing RESTful resources with the HTTP protocol, the resource URL acts as a resource identifier. The standard HTTP operations known by REST are GET, PUT, DELETE, POST, and HEAD. REST-based web service has several advantages over SOAP. For most servers, RESTful web services provide a good caching infrastructure over an HTTP GET method. It also offers an effective way of interacting with lightweight clients, such as smartphones.

Currently, REST-based web services are already implemented in various applications, especially on web services that have a major mobile device client [7]. In this research implemented RESTful web service in terms of data exchange.

Increasing the amount of data requires more storage space on the server. To save data storage on the server can be done in various ways. One way to save data storage on a server is to compress files or data before saving it on the server. One of the most popular and well-suited compression algorithms for text data compression is the LZW algorithm [8], [9]. In this research, we implement it to compress data before it saved in the database. This is because the LZW compression algorithm which is a dictionary-based compression is very effective in compressing text data that has many letters, words or repetitive sentences [10]. The more repertory of the combined strings in an additional dictionary on a plaintext the better the plaintext compression result [11]. The purpose of this implementation is to reduce the size of the storage space on the server database with the LZW compression algorithm.

The LZW algorithm forms a list of tables that will encode the sequence of symbols into the N-bit index contained in the table. Table size has 2^N dictionary list. The general principle of how the LZW algorithm works is to check each emerging character and then merge with the next character into a new string. If the

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

new string is not in the dictionary or has not been indexed then the new string will be added to the dictionary.

We also proposed encryption algorithm for securing data exchange through internet network. In addition to speed, the exchange of data through the Internet network also need to consider the safety factor. Currently, there are several commonly used encryption algorithms, including DES, 3DES, AES, Blowfish, RC4, and RSA. DES algorithm is an encryption algorithm that is quite old and has some security issues [12], [13]. The 3DES algorithm is a development of DES, but still lacks the time side of the slow process. Both algorithms are designed to be applied to hardware-based encryption. Meanwhile, the AES algorithm is a development of the DES that is set to standard by the US federal organizations. This algorithm is very secure because it uses keys 128, 192 or 256 bits.

Some studies have compared the encryption algorithm to determine the performance of each algorithm. It is quite difficult to determine the best encryption algorithm because of many factors that influence it. However, in this study, the encryption algorithm is determined based on two main criteria: speed and security. In terms of speed, most researchers who have compared the capabilities of some encryption algorithms agree to conclude that the Blowfish algorithm is the best [14]–[18]. Meanwhile, in terms of security, the best algorithm according to previous research is AES [12], [15], [19]. Therefore, this study used two algorithms in sequence: Blowfish and AES with 128-bit keys.

Blowfish encryption requires 64 bits of plaintext blocks as input and produces 64 bits of ciphertext. The key size for Blowfish can be selected in the range of 32 bits to 448 bits. The larger the bit size, the higher the level of security. The input blocks are split in half L0 and R0 where each contains 32 bits.

II. PROPOSED METHOD

A. The Application Architecture

Based on the above background, in this study, we developed a web-based data exchange model. We use the REST web service method because it has advantages in terms of speed and data security. Meanwhile, to optimize data security in web service data exchange process, we utilize the REST web service with the Blowfish and AES-128 encryption algorithm. And to improve the effectiveness of data storage on the server, we apply the LZW compression algorithm. In this study also developed a mobile application that implements the proposed model.

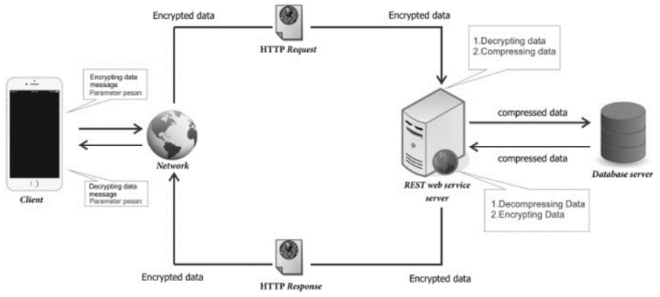


Fig. 1. The application architecture of the Proposed Model

TABLE I. WEB SERVICE DESIGN

Services	Method	Parameter	Response
Class list	GET		Class data (array)
Join Class	POST	id_kelas (integer)	Status (bool), message (string)
Unjoin Class	POST	id_kelas (integer)	Status (bool), message (string)
Class search	POST	carikelas (string)	Class data (array)
Class Delete	POST	id_kelas (integer)	Status (bool), message (string)
Create class	POST	Class data (array)	Status (bool), message (string)
Update class	POST	id_kelas (integer), Class data (array)	Status (bool), message (string)
User presence	POST	id_kelas (integer), id_user (integer)	Status (bool), message (string)
User Profile	GET	id_user (integer)	User data (array)
User followers List	GET	-	User follower data (array)
User following list	GET	-	User following data (array)
Add user skill	POST	id_user (integer), Skill data (array)	Status (bool), message (string)
Find user based-on event	POST	Keyword (string), id_event (integer)	User data (array)
Find user based-on name	POST	Keyword (string), id_event (integer)	User data (array)
Find user based-on skill	POST	Keyword (string), id_event (integer)	User data (array)
Follow an user	POST	id_user (integer), id_user2 (integer)	Status (bool), message (string)
Unfollow user	POST	id_user (integer), id_user2 (integer)	Status (bool), message (string)

Figure 1 presents the data exchange model proposed in this study. When a client requests a service, the data is encrypted using Blowfish and produce cyphertext of original data. Then the resulting ciphertext is encrypted again with the AES-128 algorithm to produce the final ciphertext. Upon receipt of the request, the web service decrypts the data and executes commands on request from the client. Similarly, when the web service sends the response data to the client, the data is encrypted first. The implementation of the encryption algorithm makes the exchange of data between the client and the web server to be more secure.

Meanwhile, the LZW compression algorithm is also applied when the application stores data to the database server. Before it is stored, the data is compressed in advance so as to save the storage capacity in the database server. In contrast, when retrieving data from a database, it is decompressed to restore the data to its original format. This data compression will save storage space on the server.

B. *Web service design*

To apply the optimized web service model with Blowfish + AES-128 encryption algorithm and the LZW compression, we proposed a REST-based web service that serves multiple requests. The service request method consists of a POST and GET. Table 1 lists the services provided by the web service.

III. Results and Discussions

This section describes the prototype of mobile-based applications that implement encryption and compression algorithms, test results against applications and discussions.

A. *The Prototype of the Proposed Method*

In this study, we developed a mobile-based application that implements the Blowfish and AES-128 encryption algorithms, as well as the LZW compression algorithm. The developed application is part of BluCampus application of Budi Luhur University. The application displays services and information related classes in the form of courses, workshops, and seminars that can be followed by students of Budi Luhur University. Each user can register and follow the ongoing activities. The activity committee can also record the attendance of the participants through its mobile device.

Figure 2 presents some of the mobile-based application views developed in this study. Figure 2 (a) shows the list of active and participatory classes. Meanwhile, in Figure 2 (b) a page to create a class or course is shown, and Figure 2 (c) displays the feature of recording attendance in a class or activity.

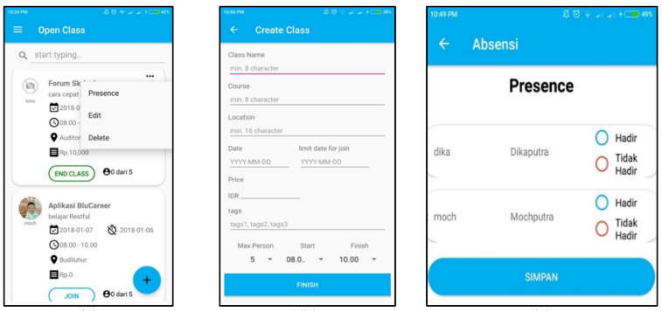


Fig. 2. Some of the application views, (a) active class list, (b) class creation form, and (c) attendance recording feature.

B. *Encryption Testing on Data Exchange*

Blowfish and AES-128 encryption algorithms are applied to the process of sending data from client to server and vice versa. Figure 3 shows an example of sending unencrypted data and Figure 4 shows an example of sending encrypted data. With the application of encryption algorithm in the data delivery process, data security will be better than sending data without encrypted.

C. *Compression Testing*

In Table 2 we present the results of compression and decompression process testing on a number of data. Test data in the form of text with a character length of 100, 200, 400, 800 and 1600. Each data is tested 10 times using the device RESTful Stress Tool and internet speed of 10 Mbps. The compression ratio (CR) and the Saving ratio (SR) are calculated with (1) and (2) [20], [21].

$$CR = \frac{uncompressed\ size}{compressed\ size} \quad (1)$$

$$SR = 1 - \frac{compressed\ size}{uncompressed\ size} \quad (2)$$

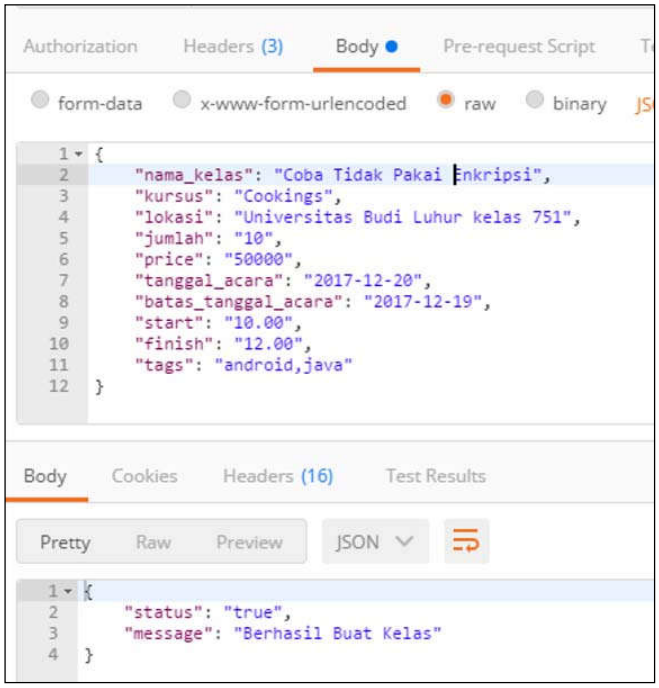


Fig. 3. Unencrypted data exchange

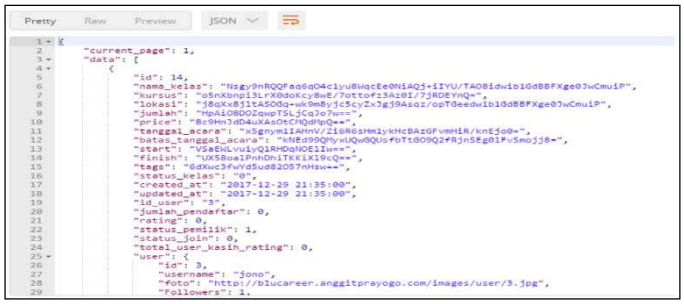


Fig. 4. Encrypted data exchange

TABLE II. The Experiment Results of the Compression Algorithm

Charac-ters	Original size (bytes)	Compressed size (bytes)	Comp. times (ms)	CR	SR	Dcomp. times (ms)
100	102	86	153	1.19	0.16	137
200	202	145	160	1.39	0.28	140
400	402	259	192	1.55	0.36	154
800	802	461	218	1.74	0.43	166
1600	1602	783	260	3.01	0.67	206
			Average	1.78	0.38	

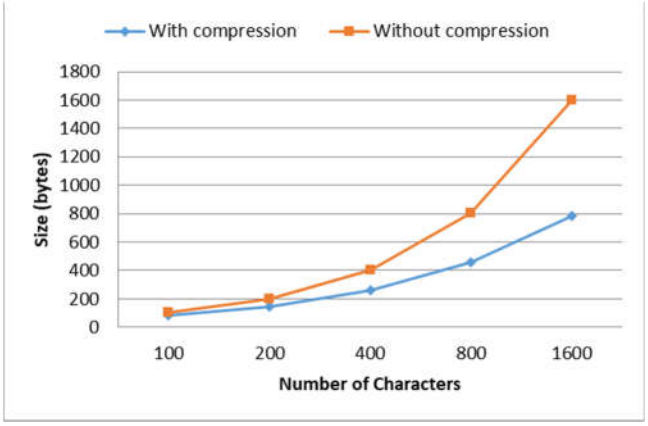


Fig. 5. Comparison of compressed file sizes with uncompressed files.

Test results showed a significant change in the size of the data after compression. The average compression rate of all tests was 1.78, and the resulting storage capacity savings were 38%. From the test results can also be concluded that the bigger the file size, the higher the compression ratio and the more storage space savings. Figure 5 presents a comparison of the size of the original data with the compressed data. In the meantime, Figure 6 shows the time comparison to store data between compressed and uncompressed data for every testing data.

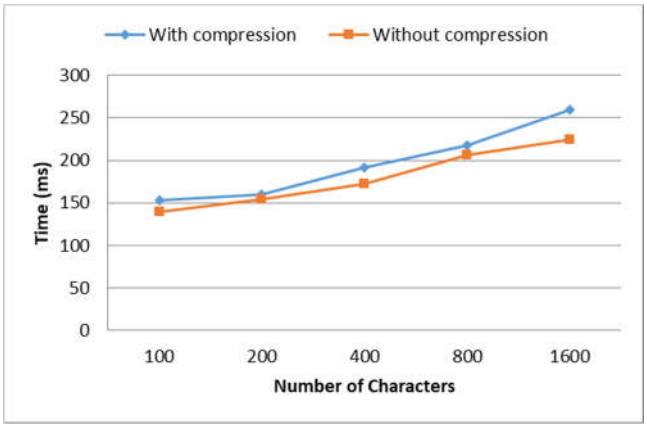


Fig. 6. Time comparison to store data between compressed and uncompressed data.

D. Web Service Testing

In this research, we also tested the web service with Blackbox testing method. Table 3 presents some of the web service services tested in this study. Based on the test results can be concluded that all services can run well.

IV. Conclusion

Based on the results of tests that have been done in this study, can be drawn some conclusions. This research produced a web service with several services. In addition, this study also produced a prototype of mobile-based applications. In testing the process of encryption of the data, all tested data successfully exchanged between web service and client. Web service testing showed that all services could run well.

TABLE III. The testing results of the web service experiments

Num	Services	Results
1	List of the class	Success
2	Create new class	Success
3	Details of the class	Success
4	Update class	Success
5	Delete class	Success
6	Find class	Success
7	Join class	Success
8	Unjoin class	Success
9	List of joined class	Success
10	Class presence	Success
11	End the class	Success
12	User profile data	Success
13	User feedback data	Success
14	Create feedback	Success
15	User skill data	Success
16	Another user skill user	Success
17	Add new skill user	Success
18	Add user about me	Success
19	Delete skill	Success
20	List of the class of a user	Success
21	Class rating of the user.	Success
22	Class data from another user	Success
23	Class data of the joined users	Success
24	Class data of the another joined-user	Success
25	Seminar data of the joined user	Success
26	Seminar data of the another joined-user	Success
27	Workshop data of the joined user	Success
28	Workshop data of the another joined-user	Success
29	Find user data based on event	Success
30	Find user data based on the name	Success
31	Find user data based on skill	Success
32	List of the user followers	Success
33	List of the user following	Success
34	List of the followers of another user.	Success
35	List of the following of another user.	Success
36	Follow a user	Success
37	Unfollow a user	Success
38	List of user conversations	Success
39	The message of the user data	Success
40	Create message	Success

In the test of the compression algorithm, it is concluded that the average compression ratio of 1.78. By compressing the data, the average savings of storage space on the server as much as 38%. The application of compression algorithm has an impact on the increase of computation time, although not significant. It shows that the compression algorithm can optimize the implementation of web service.

References

[1] Statista, "Number of smartphone users worldwide from 2014 to 2020 (in billions)," *Statista.Com*, 2018. [Online]. Available: https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/. [Accessed: 01-Apr-2018].

[2] StatCounter, "Desktop vs Mobile vs Tablet Market Share Worldwide," *StatCounter.Com*, 2018. [Online]. Available: http://gs.statcounter.com/platform-market-share/desktop-mobile-tablet. [Accessed: 02-Apr-2018].

[3] M. Zur Muehlen, J. V. Nickerson, and K. D. Swenson, "Developing web services choreography standards - The case of REST vs. SOAP," *Decis. Support Syst.*, vol. 40, no. 1 SPEC. ISS., pp. 9–29, 2005.

[4] I. Al Rassan and H. Alyahya, "Tagged-sub Optimal Code (TSC) compression for improving performance of web services," *Procedia*

Comput. Sci., vol. 62, no. December 2015, pp. 167–169, 2015.

[5] A. R. Shukla and P. P. D. Thakare, "A Review On Performance Enhancement Of Web Services Using Tagged-Sub Optimal Code (TSC) Compression," *Int. Res. J. Eng. Technol.*, vol. 5, no. 3, pp. 1177–1180, 2018.

[6] A. M. Abbas, A. A. Bakar, and M. Z. Ahmad, "Fast dynamic clustering SOAP messages based compression and aggregation model for enhanced performance of Web services," *J. Netw. Comput. Appl.*, vol. 41, no. 1, pp. 80–88, 2014.

[7] H. Hamad, M. Saad, and R. Abed, "Performance Evaluation of RESTful Web Services for Mobile Devices.," *Int. Arab J. e-Technol.*, vol. 1, no. 3, pp. 72–78, 2010.

[8] T. A. Welch, "A Technique for High-Performance Data Compression," *IEEE Comput.*, pp. 8–19, 1984.

[9] J. Ziv and A. Lempel, "Compression of Individual Sequences via Variable-Rate Coding," *IEEE Trans. Inf. Theory*, vol. 24, no. 5, pp. 530–536, 1978.

[10] R. Maulunida and A. Solichin, "Optimization of LZW Compression Algorithm With Modification of Dictionary Formation," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 12, no. 1, p. 73, 2018.

[11] Hidayat, W. Zarman, and T. Pamungkas, "Implementasi Algoritma Kompresi Lzw Pada Database Server," *J. Ilm. Komput. dan Inform. (KOMPUTA)*, vol. 2, no. 1, pp. 7–13, 2013.

[12] T. Pornin, "Comparison of DES, Triple DES, AES, blowfish encryption for data," *StackOverflow.Com*, 2011. [Online]. Available: https://stackoverflow.com/questions/5554526/comparison-of-des-triple-des-aes-blowfish-encryption-for-data. [Accessed: 24-Apr-2018].

[13] A. Solichin and E. W. Ramadhan, "Enhancing data security using DES-based cryptography and DCT-based steganography," in *2017 3rd International Conference on Science in Information Technology (ICSITech)*, 2017, pp. 618–621.

[14] J. Thakur and N. Kumar, "DES, AES and Blowfish: Symmetric key cryptography algorithms simulation based performance analysis," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 1, no. 2, pp. 6–12, 2011.

[15] P. Patil, P. Narayankar, D. G. Narayan, and S. M. Meena, "A Comprehensive Evaluation of Cryptographic Algorithms: DES, 3DES, AES, RSA and Blowfish," *Procedia Comput. Sci.*, vol. 78, no. December 2015, pp. 617–624, 2016.

[16] S. P. Singh and R. Maini, "Comparison of Data Encryption Algorithms," *Int. J. Comput. Sci. Commun.*, vol. 2, no. 1, pp. 125–127, 2011.

[17] A. Nadeem and M. Y. Javed, "A Performance Comparison of Data Encryption Algorithms," in *2005 International Conference on Information and Communication Technologies*, 2005, pp. 84–89.

[18] D. S. A. Elminaam, H. M. A. Kader, and M. M. Hadhoud, "Performance Evaluation of Symmetric Encryption Algorithms," *IJCSNS Int. J. Comput. Sci. Netw. Secur.*, vol. 8, no. 12, pp. 280–286, 2008.

[19] P. Princy, "A Comparison of Symmetric Key Algorithms DES , AES , Blowfish, RC4, RC6: A Survey," *Int. J. Comput. Sci. Eng. Technol.*, vol. 6, no. 5, pp. 328–331, 2015.

[20] Wikipedia, "Data compression ratio," *Wikipedia.com*, 2017. [Online]. Available: https://en.wikipedia.org/wiki/Data_compression_ratio#cite_note-CompressionRatioBroadcastEngineering-1. [Accessed: 10-Mar-2018].

[21] C. Poynton, *Digital Video and HD: Algorithms and Interface*, Second Edi., no. November. Elsevier Inc., 2012.

Performance Discovery in Ubiquitous Networks

Irawan Dwi Wahyono, Agusta Rakhmat Taufani, Irham Fadlika, Hakkun Elmunsyah, Gradiyanto Radityo Kusumo
Department of Electrical Engineering
Universitas Negeri Malang
Malang, Indonesia
Email: irawan.dwi.ft@um.ac.id, agusta.rakhmat.ft@um.ac.id, irham.fadlika.ft@um.ac.id, hakkun@um.ac.id, gradiyanto.radityo.1405334@students.um.ac.id

***Abstract*— Optimization discovery is an ongoing study on Ubiquitous Networking due to some limitations and constraints in the device or node, i.e., calculation, limited power, bandwidth, high mobility at each node. In this research developed Discovery protocol in the Ubiquitous network where service descriptor optimization package for service request and service advertisement with the classification of tree model/taxonomy. Distribution of packets on the network layer by utilizing the routing protocol OLSR. The method of developing the protocol of this invention in this study is called M-UBI. In this research, the result of test and data analysis with simulation, the success of M-UBI in overcoming bandwidth and delay constraint in Ubiquitous is the decrease of bandwidth equal to 31,6% and a decrease of delay equal to 16,7%.**

Keywords—Ubiquitous, Discovery, OLSR

I. INTRODUCTION

Ubiquitous applications have widely used in human and environmental interactions by utilizing wifi, 3G, WiMax and Bluetooth networks [1-2]. Devices used in Ubiquitous are mobile. The protocols used in mobile devices are similar to those used in the ad hoc mobile protocol [3-5]. The problem with mobile devices is in searching for services that are owned by devices everywhere in the ubiquitous network [6-9]. Another problem is the quality of service on the network in Ubiquitous [10-12]. Research on ubiquitous and mobile ad hoc has widely used in the comparison protocols used between AODV and OLSR by looking at the performance of the routing protocol [13-14]. Moreover, also performance in resource allocation scheme in Ubiquitous Network to get quality service on the Ubiquitous network [15-17].

In Ubiquitous, it usually consists of mobile users with different functions and uses, embedded systems, monitoring applications, control applications with multiple sensors and actuators, and resources shared with users.[10-12]. There are various ways to resolve ubiquitous service complexity by providing all the resources of the ubiquitous service and the ownership of the service discovery function on each node on the ubiquitous network. There are architectural methods, and discovery frameworks that allow devices/nodes to find and retrieve services on the MANET network, as well as existing resources, advertise on mobile devices/nodes in Ubiquitous [10-12].

Inside the network can occur without forcing the user to enter an IP address, password, username or other attribute value. This method is Konark [13], Konark is implemented on the application layer protocol side regardless of the packet carrier for distribution at the routing protocol layer. The disadvantage of the Konark method and framework is the length of the data packets in defining the descriptors. The longer distributed data packets cause overhead across networks everywhere [14-15]. The cause of overhead everywhere can also be due to the selection of routing protocol methods for the distribution of data packets as operators. Routing protocols are of two kinds: reactive and proactive. Therefore, the selection of essential methods in the optimization.

The existing methods have a long data packet in the detection of application layer protocols. It is, therefore, necessary to reduce the data packet short and also to determine the routing protocol of the desired operator on the network protocol layer. In each communication layer have their respective protocols, to connect between two layers of the protocol that is between two application layer protocols and network layer is also called cross-layer protocol [16-17].

In this research developed discovery method in the ubiquitous network by modifying data packet on OLSR. The method used involves the discovery of a taxonomic tree in the protocol routing format in OLSR. Usage, The use of taxonomy, makes it easier for discovery services on ubiquitous networks. Disseminate discovery in advertisements and requests on the ubiquitous network using the Dynamic Bloom filter and performing local cache updates on each ubiquitous device. The method of tree taxonomy and discovery service on the ubiquitous network is M-UBI.

II. DISCOVERY DESIGN

A. Taxonomy Tree.

Classification of information services in the form of trees to facilitate requests for desired discovery information and can perform advertising messages for discovery [17-19]. An example of making a tree-shaped classification for a discovery device or resource contained in a ubiquitous network shown in Figure 1. The discovery process is first performed at the highest root and then down to the specification of each service information contained in each node. The root service in Figure 1. is the highest root in determining the service path to be searched while on the path implement must complete and detail based on Discovery classification in Figure 1.

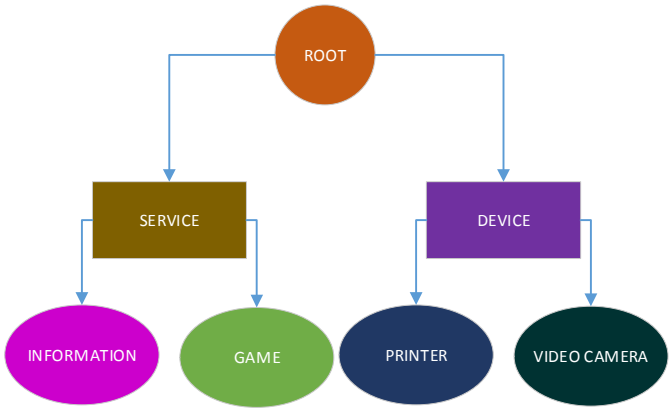


Fig. 1. taxonomy tree

B. *Overview Design*

This system design designed for discovery of limited bandwidth and delay environments, to optimize discovery. This research proposes discovery with M-UBI method. The working principle of M-UBI is to describe the description discovery efficiently by utilizing dynamic bloom filters that are classified in the form of trees and deploy services by flooding service information on OLSR routing messages and utilizing cache on local devices by way of service advertisement as shown in Figure 2. The location of the discovery in the ubiquitous network protocol is shown in Figure 3 where M-UBI set in the application layer by utilizing the OLSR protocol routing format on frees data space distributed on both UDP and TCP transport layers via the Hello packet.

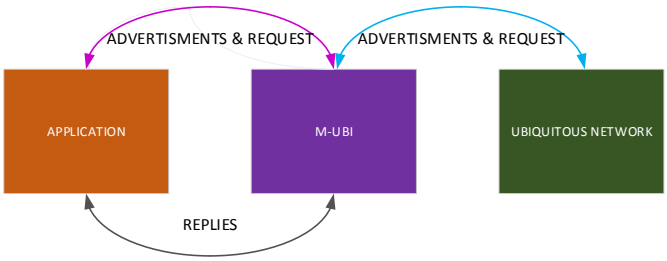


Fig. 2. Design system

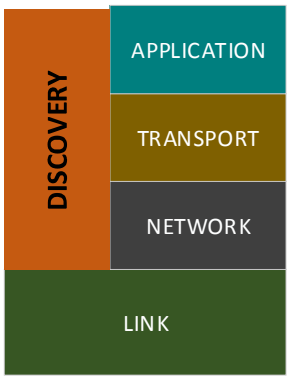


Fig. 3. Cross-layer in the system

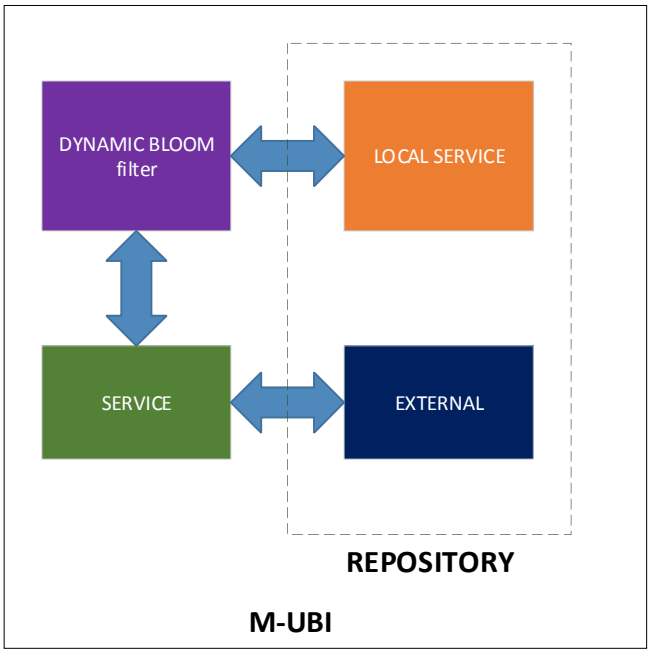


Fig. 4. Working principle of M-UBI

The working principle of M-UBI is shown in Figure 4 where M-UBI interacts through the OLSR protocol routing network as the OLSR packet carrier functions to broadcast packets by utilizing MPR flooding techniques. Package description using Dynamic Bloom Filter [17-19]. In M-UBI there are four main components:

1. Dynamic Bloom filter (DBf) functions as the conversion of all data discovery.

2. Service serves as addressing service requests and discovery ads from external nodes and local nodes.

3. Externally have as Message Creator function to create new message before broadcast Ubiquitous network that is Service Advertisement (SA) and Service Request (SR)

4. Repository serves as data storage service.

The perform by M-UBI is shown in figure 5 is as follows: If the node wants discovery of the S device described in text form, then the node makes a service request by utilizing DBf by changing the discovery description into a small packet in DBf form. The service request form DBf has become B (SR). B (SR) then performs a query discovery contained in the External Service repositories, if the discovery is not found then the node creates a new message request again containing all the service requests that are in the service request repository into a single service request by utilizing DBf to B (SR) then the node sends service through ubiquitous network by utilizing OLSR routing protocol with MPR flooding technique.

If the node receives B (SR) from the external node, then it is queried with its repository which lists the discovery on the local node. If discovery found, the node will respond immediately by doing the Advertisement service. Service Advertisement is

created using DBf which consists of all services contained in the Won Service repository into B (SA) and then sent to the ubiquitous network using OLSR routing protocols with MPR flooding techniques.

If the node wants to do the device advertisement, the S-description tool immediately added to the Local Service Repository, then create an advertisement service that consists of all services contained in the Local Service repository by utilizing DBf to B (SA) and then sent to the Ubiquitous network. When the node receives service advertisement B (SA) from the network, the node immediately adds, updates and queries to External Service repositories.

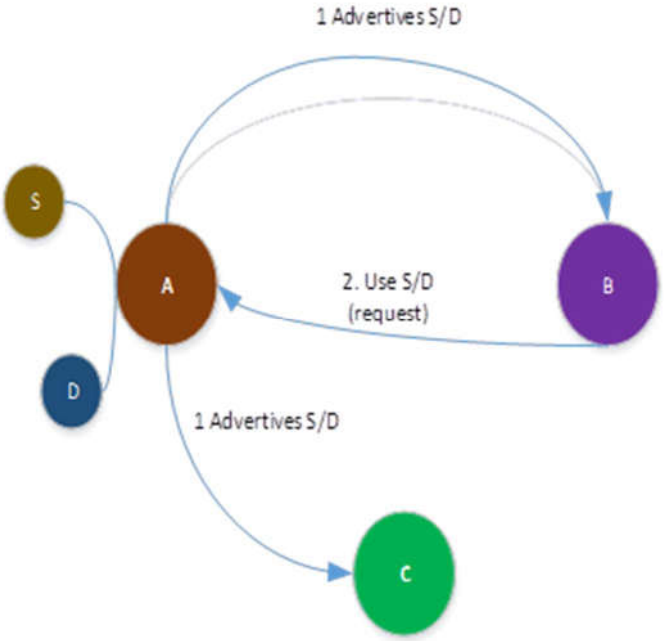


Fig. 5. The perform by M-UBI

C. Service description

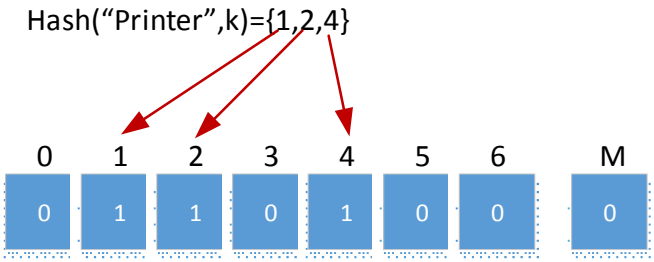


Fig. 6. The first discovery added to the DBf

The Dynamic Bloom filters (DBf) technique used initially as an application in the database, but due to DBf characteristics, this technique becomes a reference in various research, especially research computer network either ad-hoc communication or infrastructure in route determination package, determination of size of data, data security, data encryption, by using Dynamic Bloom filter. [17-19]. The use of DBf considered in every application where the use of space lists is significant, and a small number of false positives are acceptable. The algorithm and method of Dynamic Bloom filter is the result of development on Bloom filter method.

In this research using Dynamic Bloom filter where DBf has an efficient way to describe device/service by using DBf. DBf accommodates all the resources in the form of device/service available on a single ubiquitous node or a particular directory in small array sizes as shown in Figure 6. This approach provides efficient and timely deployment of services in ubiquitous networks.

In ubiquitous networks with sizeable available service scopes, it will cause false positives in service discovery in ubiquitous device. false positives in service discovery occur when the device / node responds to a positive request but the requested service / device does not exist as shown in Figure 7 where the device or service is not networked. To reduce false positives can be reduced by analyzing the DBf data format and then adjusting the data format to DBf parameters with different DBf values to obtain a small false positive. DBF declaration in the form of array pointer in the program list shown in Figure 8.

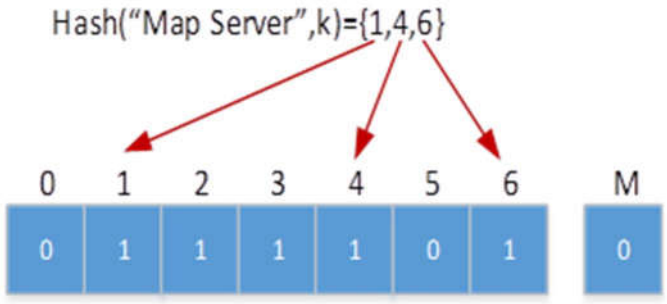


Fig. 7. This query yields a false positive

```
///Pointer pada  local Dynamic Bloom filters
(DBf) (
bitset<DBf_SIZE>* ::getLocalDBf()
{
        return &_localDBf;
}
```

Fig. 8. Declare DBf in the array pointer

D. False Positive Dynamic Bloom filter

The parameter in false positive is m that is the bit length DBf, n is expressed as the number of service / device descriptions included in the DBf, and k is expressed as a hash function used, The false positive rate is formulated in equation 1 [39].

$$P_{\text{fp}} = (1 - e^{-\frac{kn}{m}})^k \quad (1)$$

In equation 1 it is shown that the total service is a variable value parameter when the discovery service process on the target discovery service. Therefore, selection of k and m values can decrease the occurrence of false positive requests.

In this DBf there are two ways to minimize false positives [17-19]:

1. Change the sum of k hash functions.
2. Raise the format of the DBf itself,

The determination of the best k value can be solved by taking the derivative from equation 1 and then determining that the total hash function. While to determine the width of DBF is m and determine the length of service description that is n can be completed with equation 2:

$$k = \frac{m}{n} = \ln 2 \Rightarrow k_{\text{opt}} = [k] \qquad (2)$$

In designing the discovery protocol based on DBf, based on equation 2 it is essential to the selection of an optimal hash function value. The parameter is to obtain the number of services that stored in ubiquitous nodes and DBf width which has limitations due to communication protocols using wireless. In M-UBI, the default value of the hash function is 4.

While in designing a minimal false positive level, mathematical calculations for the m filter should be as vast as possible if possible, unlimited. However, due to computational time, network data speed and memory usage limit the default filter size. M-UBI using default filter size is 128 bit

E. Protocol Format

DBf is used to describe the service request and advertised requests. Discovery in M-UBI performs distributed DBf using OLSR routing protocol format [44] in free space as shown in Figure 8. Process in OLSR for default Control messages that are HELLO packets and TC packets. The packet was communicated using an integrated package format in OLSR [22]. When broadcasting an OLSR message packet, it can consist of several messages using the piggyback method on OLSR with the M-UBI message format as described in Figure 9.

In OLSR serve a way to distribute different types of data through an intermediate node. An intermediate node can also facilitate if the node does not support specific messages due to different messages, but will continue using the default-forwarding algorithm. In messages to be distributed using OLSR routing protocols have own headers, and the sender can restrict the transmission flooding for multiple hops or increase a specific official time for a particular message.

M-UBI integrated with OLSR standard as shown in Figure 9 in which the M-UBI message format will be sent as part of the data from the OSLR public message format with message types that can set according to the valid time. Time to Live is set to 255.

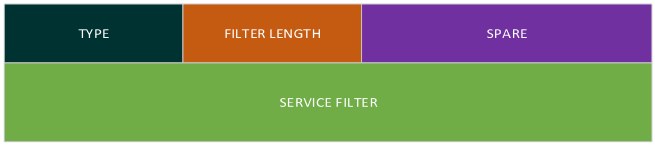


Fig. 9. M-UBI format

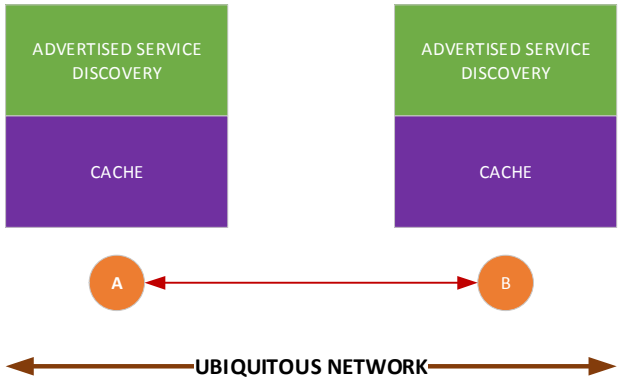


Fig. 10. Communication on the Ubiquitous Network

III. IMPLEMENTATION

The implementation in this study used an NS-2 simulation [21] where routing protocols use OLSR [22] then in advertisement and request on the ubiquitous network using M-UBI protocol as shown in Figure 11. Service information or service description placed on the expressed in spare data in the OLSR package format as described in Figure 12.

The service/device description declaration is shown in Figure 13 which uses hash functions in the Dynamic Bloom filter. In this research, M-UBI Design utilizes MD5 [39] algorithm in the following way: The determination of the value of the hash function in BBf was developed by an MD5 algorithm by determining the value of k on each bit result with a hash length: 128-bit. On each sub-section on MD5 as input on the function. Each k function is set to one bit in the BBf v as shown in Figure 14.

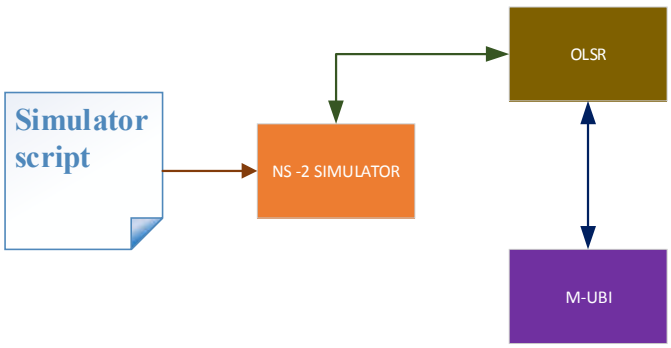


Fig. 11. Implementation in NS2

```
struct m-ubimsg
{
olsr_u8_t type;
olsr_u8_t length;
olsr_u16_t spare;
unsigned char filter[FILTER_LENGTH];
};
```

Fig. 12. OLSR Structure

```
string ::getmd5Hash(string* service)
{
      MD5 context;
      context.update((unsigned char*)service->c_str(),service-
>size());
      context.finalize();
      char* digest = context.hex_digest();

      string line = digest;
      return line;
}

unsigned long int ::ConvertHash2Int(string hash)
{
      std::stringstream stream;
      stream << hash;
      unsigned long int n;
      stream >> std::hex >> n;
      return n;
}
```

Fig. 13. Hash function di M-UBI

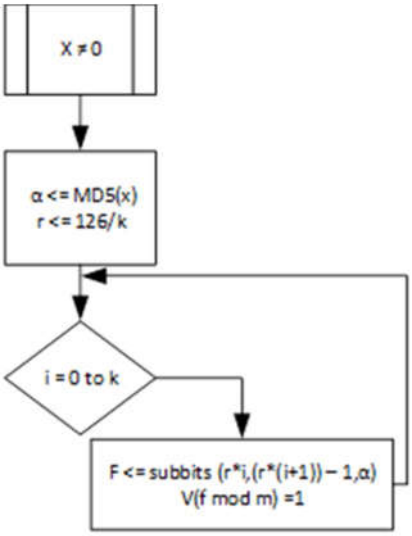


Fig. 14. Determination of Value k

IV. PERFORMANCE EVALUATION

A. Simulation Setup

In testing the design can be done using one of the methods in previous research on MANET service discovery in which this testing approach to take different benefits and losses into consideration of previous research. The choice of test method depends on the parameters used, the resources used for both hardware and software, routing protocols used and the selection of the discovery framework used. In the previous study, service discovery studies only evaluated in an analytical or framework manner [20,23], and other descriptions are purely evaluated by simulations [20,23] and also while others implemented in real-world experiments in small scopes.

Previous studies of service discovery only test different scenario models in simulations and experiments. Therefore, this study is challenging to compare several methods of research results. The choice of scenarios as differences in targeted service discovery solutions for different applications and one evaluation of methods or research results do not apply to all solutions. Various configurations found in different implementations in ubiquitous tissue then a standard set of scenarios was required to evaluate different discovery studies [20,23]. However, service discovery testing scenarios for various purposes, all service discovery scenarios depend on the simulation as a ratification method.

In this research, several approaches in the simulation of M-UBI implemented in NS-2.35 Network Simulator and plug-in OLSR [22]. The M-UBI configuration of NS-2.35 discribed in Table 1.

TABLE I. PARAMETER

Parameter	Value
Application Simulator	NS-2.31
OS	Ubuntu
TR (Transmission Range)	100m
MAC Address	802.11
RF (Reflection)	Two Ray Ground
MV (Movement)	Random Waypoint
RP (Routing)	OLSR
Settings OLSR	Default

B. *Bandwidth*

Bandwidth is the number of bytes on the node required during the service discovery process and tested using different network topologies. The bandwidth testing of discovery on the ubiquitous network using M-UBI and Non-M-UBI using OLSR routing protocol obtains trace piece result in NS-2 as shown in Figure 15. The test is done by the number of nodes 4,9,16,15, 36, 39 and 64. The result of using bandwidth on the number of nodes shown in Table 2. The result of this bandwidth measurement obtained efficiency with the use of M-UBI of 31.6% as shown in the graph in Figure 16.

```
s 0.000169789 _0_ RTR  --- 0 OLSR 48 [0 0 0 0] ------- [0:255 -1:255 32 0] [1 0 [HELLO 0 0 0]]
r 0.001478055 _1_ RTR  --- 0 OLSR 48 [0 ffffffff 0 800] ------- [0:255 -1:255 32 0] [1 0 [HELLO 0 0 0]]
s 0.044061336 _1_ RTR  --- 1 OLSR 48 [0 0 0 0] ------- [1:255 -1:255 32 0] [1 0 [HELLO 1 0 0]]
r 0.045249602 _0_ RTR  --- 1 OLSR 48 [0 ffffffff 1 800] ------- [1:255 -1:255 32 0] [1 0 [HELLO 1 0 0]]
r 0.045249602 _2_ RTR  --- 1 OLSR 48 [0 ffffffff 1 800] ------- [1:255 -1:255 32 0] [1 0 [HELLO 1 0 0]]
s 0.156480886 _2_ RTR  --- 2 OLSR 48 [0 0 0 0] ------- [2:255 -1:255 32 0] [1 0 [HELLO 2 0 0]]
r 0.157969152 _1_ RTR  --- 2 OLSR 48 [0 ffffffff 2 800] ------- [2:255 -1:255 32 0] [1 0 [HELLO 2 0 0]]
```

Fig. 15. Example of a packet tracer in NS2

TABLE II. BANDWIDTH MEASUREMENT RESULT

Byte	**number of nodes**						
	4	9	16	25	36	49	64
M-UBI	0	58	102	204	344	554	733
NON M-UBI	172	366	512	722	1170	1420	1942

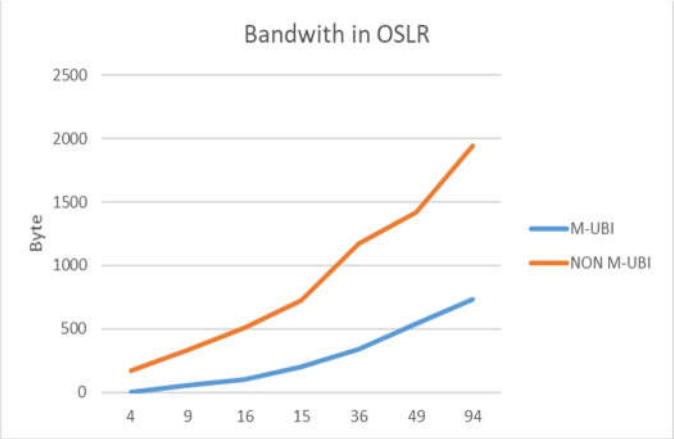


Fig. 16. Graph of Measurement Bandwidth

C. Delay

Delay is a delay evaluation (time consumed) to perform discovery on network ubiquitous. Test of delay discovery in a ubiquitous network using M-UBI and Non-M-UBI using OLSR routing protocol. The result is a trace piece in the NS-2 as shown in Figure 17. The test is carried out by the number of nodes 4,9,16,15,36,39 and 64. The result of the delay in the HELLO data packet with the number of nodes shown in Table 3. Results from this delay measurement obtained a decreased efficiency of 16.7% if using M-UBI as shown in the graph in Figure 18.

```
s 0.044061336 _1_ RTR  --- 1 OLSR 48 [0 0 0 0] ------- [1:255 -1:255 32 0] [1 0 [HELLO 1 0 0]]
r 0.045170002 _2_ RTR  --- 1 OLSR 48 [0 ffffffff 1 800] ------- [1:255 -1:255 32 0] [1 0 [HELLO 1 0 0]]
r 0.045170043 _3_ RTR  --- 1 OLSR 48 [0 ffffffff 1 800] ------- [1:255 -1:255 32 0] [1 0 [HELLO 1 0 0]]
r 0.045170043 _0_ RTR  --- 1 OLSR 48 [0 ffffffff 1 800] ------- [1:255 -1:255 32 0] [1 0 [HELLO 1 0 0]]
s 0.082072703 _4_ RTR  --- 2 OLSR 48 [0 0 0 0] ------- [4:255 -1:255 32 0] [1 0 [HELLO 4 0 0]]
r 0.083361410 _3_ RTR  --- 2 OLSR 48 [0 ffffffff 4 800] ------- [4:255 -1:255 32 0] [1 0 [HELLO 4 0 0]]
r 0.083361410 _0_ RTR  --- 2 OLSR 48 [0 ffffffff 4 800] ------- [4:255 -1:255 32 0] [1 0 [HELLO 4 0 0]]
s 0.156480886 _2_ RTR  --- 3 OLSR 48 [0 0 0 0] ------- [2:255 -1:255 32 0] [1 0 [HELLO 2 0 0]]
r 0.157629552 _1_ RTR  --- 3 OLSR 48 [0 ffffffff 2 800] ------- [2:255 -1:255 32 0] [1 0 [HELLO 2 0 0]]
s 0.319432572 _3_ RTR  --- 4 OLSR 48 [0 0 0 0] ------- [3:255 -1:255 32 0] [1 0 [HELLO 3 0 0]]
r 0.320401279 _1_ RTR  --- 4 OLSR 48 [0 ffffffff 3 800] ------- [3:255 -1:255 32 0] [1 0 [HELLO 3 0 0]]
r 0.320401279 _4_ RTR  --- 4 OLSR 48 [0 ffffffff 3 800] ------- [3:255 -1:255 32 0] [1 0 [HELLO 3 0 0]]
```

Fig. 17. Example of a packet tracer in NS2

TABLE III. RESULT MEASUREMENT DELAY

Delay(s)	number of nodes						
	4	9	16	25	36	49	64
M-UBI	1	2,2	3,9	7,8	8,58	10,3	20,6
NON M-UBI	1,2	2,64	4,68	9,36	10,3	12,4	24,7

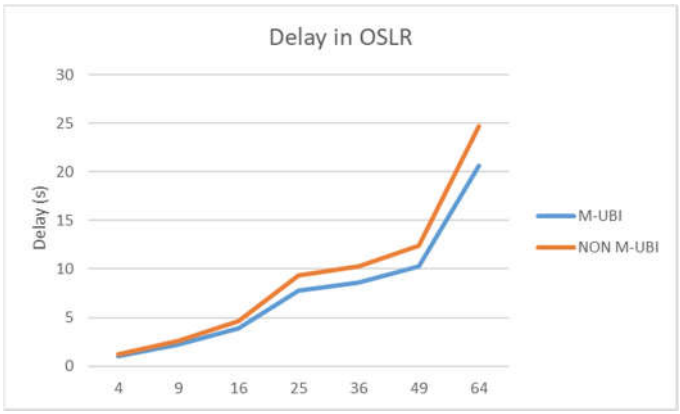


Fig. 18. Graph of Measurement Delay

V. CONCLUSION

In this research, we developed a discovery method for services and devices on the ubiquitous network called M-UBI which utilizes the protocol routing in OLSR in doing advertisement and request on the ubiquitous network. The discovery declarations on services and devices take advantage of the Dynamic Bloom filter (DBf). Test results on the M-UBI method obtained 31.6% bandwidth efficiency and delayed efficiency of 16.7 %.

ACKNOWLEDGMENT

This research supported by the PNPB Grant from Universitas Negeri Malang

REFERENCES

[1] Prekop, P., & Burnett, M. (2003). Activities, context and ubiquitous computing. *Computer communications*, *26*(11), 1168-1176.

[2] Moran, T. P., & Dourish, P. (2001). Introduction to this special issue on context-aware computing. *Human-computer interaction*, *16*(2), 87-95.

[3] Bang, A. O., & Ramteke, P. L. (2013). Manet: history, challenges, and applications. *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, *2*(9), 249-251..

[4] Macker, J. (1999). Mobile ad hoc networking (MANET): Routing protocol performance issues and evaluation considerations.

[5] Gochhayat, S. P., & Venkataram, P. (2013). Performance Analysis of a Priority based Resource Allocation Scheme in Ubiquitous Networks. *JoWUA*, *4*(3), 23-40.

[6] Wahyono, I. D. (2016). SERVICE DISCOVERY BERBASIS BREADTH BLOOM FILTER DI MOBILE AD-HOC NETWORK (MANET). *TEKNO*, *23*(1)

[7] Lestari, D., Wahyono, I. D., & Fadlika, I. (2017, November). IoT based Electrical Energy Consumption Monitoring System Prototype: Case study in G4 Building Universitas Negeri Malang. In *Sustainable Information Engineering and Technology (SIET), 2017 International Conference on* (pp. 342-347). IEEE.

[8] Kim, J. Y., & Lee, G. M. (2014, January). Context awareness for smart ubiquitous networks. In *Electronics, Information and Communications (ICEIC), 2014 International Conference on* (pp. 1-2). IEEE.

[9] Gochhayat, S. P., & Pallapa, V. (2015). An Efficient QoS Support for Ubiquitous Networks. *IEEE Transactions on Emerging Topics in Computing*, *3*(4), 524-533.

[10] Wirawan, I. M., Taufani, A. R., Wahyono, I. D., & Fadlika, I. (2017, October). Online judging system for programming contest using UM framework. In *Information Technology, Computer, and Electrical Engineering (ICITACEE), 2017 4th International Conference on* (pp. 230-234). IEEE.

[11] S. P. Gochhayat and P. Venkataram, ''Performance analysis of a priority based resource allocation scheme in ubiquitous networks,'' J. Wireless Mobile Netw., Ubiquitous Comput., Dependable Appl., vol. 4, no. 3, pp. 23–40, 2013

[12] Suwasono, S., Prihanto, D., Wahyono, I. D., & Nafalski, A. (2017). Virtual Laboratory for Line Follower Robot Competition. *International Journal of Electrical and Computer Engineering (IJECE)*, *7*(4), 2253-2260.

[13] Helal, S., Desai, N., Verma, V., & Lee, C. (2003, March). Konark-a service discovery and delivery protocol for ad-hoc networks. In *Wireless Communications and Networking, 2003. WCNC 2003. 2003 IEEE* (Vol. 3, pp. 2107-2113). IEEE

[14] Kawadia, V., & Kumar, P. R. (2005). A cautionary perspective on cross-layer design. *IEEE Wireless Communications*, *12*(1), 3-11.

[15] Stine, J. A. (2006, October). Cross-layer design of manets: The only option. In *Military Communications Conference, 2006. MILCOM 2006. IEEE* (pp. 1-7). IEEE.

[16] Skouteli, C., Samaras, G., & Pitoura, E. (2005, May). Concept-based discovery of mobile services. In *Proceedings of the 6th international conference on Mobile data management* (pp. 257-261). ACM.

[17] Bloom, B. H. (1970). Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, *13*(7), 422-426.

[18] Song, H., Dharmapurikar, S., Turner, J., & Lockwood, J. (2005). Fast hash table lookup using extended bloom filter: an aid to network processing. *ACM SIGCOMM Computer Communication Review*, *35*(4), 181-192.

[19] Rivest, R. (1992). The MD5 Message-Digest Algorithm. RFC 1321 (Informational)

[20] W. Kiess and M. Mauve. A survey on real-world implementations of mobile ad-hoc networks. *Ad Hoc Netw.*, 5(3):324–339, April 2007

[21] University of California. ns2 Network Simulator. (http://www.isi.edu.nsnam/ns/), Accessed 2008

[22] Clausen, T., & Jacquet, P. (2003). *Optimized link state routing protocol (OLSR)* (No. RFC 3626).

[23] Haerri, J., Filali, F., & Bonnet, C. (2006, June). Performance comparison of AODV and OLSR in VANETs urban environments under realistic mobility patterns. In *Proceedings of the 5th IFIP mediterranean ad-hoc networking workshop* (No. i, pp. 14-17)..

Design and Development of REST-based Instagram Spam Detector for Indonesian Language

Antonius Rachmat Chrismanto, Willy Sudiarto Raharjo, Yuan Lukito
Duta Wacana Christian University: Department of Informatics, Faculty of Information Technology
Yogyakarta, Indonesia
anton@ti.ukdw.ac.id, willysr@ti.ukdw.ac.id, yuanlukito@ti.ukdw.ac.id

***Abstract*—Instagram, the most popular picture-based social media, has been well known for spam comments, mostly for public figures, including Indonesian actors and actress. Spam comments may lead to misleading information and confusing to other followers. Our previous research on Indonesian language spam comment detector has concluded that K-Nearest Neighbor yields the best result to detect Instagram spam comments. We continue previous our work by designing the REST-based web service to detect Indonesian language spam comments for Instagram. We are using *supervised learning* combined with Instagram dataset gathered from previous work. This system is built on top of Amazon Web Service (AWS) to provide a robust, resilience, and scalable system using Java Jersey library. The result of this work is a preliminary design and implementation of REST-based Instagram spam detector for Indonesian language that has an average response time of 1678.133 ms with deviation standard of 1178.59 ms. The current service still has limitations in terms of number of features and k-d tree data structure and it still needs some optimization before it can be used in production environment.**

Keywords: REST web services, Instagram, spam comments, Indonesian Language

I. INTRODUCTION

Instagram is a very popular picture-based social media service among Indonesian actors and actress. A lot of them used Instagram to boost their popularity and to share their activities with their fans. By following their Instagram accounts, their fans are able to follow their idol visually and interact with fans. Unfortunately, it came at a cost that more and more spam comments appearing as their popularity increased. Spam comments can be a source of information misleading [1] as people might have thought that a post is full of comments, but in reality, most of them are spam. Spam also makes it harder for readers to track information in a post and also causing information becomes irrelevant against the original post by the author. Currently, there is no automatic spam remover provided by Instagram, leaving the user to clean them manually or leave them be polluting their postings.

Eskelinen [2] and Tamir [3] concluded that Instagram users need to spend extra effort to manually verify and remove all spam comments. While Instagram already provided a feature to report spam, it is not the best solution as it still requires users to remove them manually. Another possible solution is to set the account as private, but it will not be beneficial for them as it would decrease their exposure to the public. The last solution was to enable Instagram feature to remove comments that contain user-provided words that are considered as spam. None of the viable solution provided can be done automatically and works for the Indonesian language specifically.

This work is the continuation of our previous 3 years roadmap. In our previous work [4], we have gathered posting and comments data from Indonesian actor/actress with more than 10 million followers and we have collected around 17.000 data. We used the posting and comments as training data in a supervised learning system to detect Indonesian spam comments. We concluded that the best algorithm to be used for spam detection (sorted by their relevance) are K-Nearest Neighbor (k-NN) [1], Support Vector Machine [5], and Naïve Bayes [4]. We continued our work onto the next step, which is designing a service that implements the best algorithm in form of a web service. This work will be a bridge for our final goal which is to make an applied product in form of a browser plugin that can be used by all Instagram users.

II. THEORETICAL REVIEWS

A. Previous Works

Research on spam detection has been conducted in many areas. Some worked on a specific platform, such as Twitter [6] [7] [8] while others are working on various social media networks [9] [10] [11]. We are focusing on text-based spam detection that appears on comments. Spam may appear as hyperlinks, trackbacks, and pingback spam in people's posting [12]. Hines described several ways to detect spam comments by looking at the relevance, posting specific comments, links provided, the validity of the name and email address used, and whether multiple email address are used. Several techniques to reduce the number of spam is to use registration before posting any comments, using CAPTCHA, not allowing HTML code, IP address blacklisting, and to limit repetitive comment in a certain period of time [13].

In our previous work, we conducted an experiment with three different methods against a dataset of 10 Indonesian actress that has more than 10 million followers. We tested Naïve Bayes, Support Vector Machine (SVM), and k-Nearest Neighbor (k-NN) and achieve an accuracy of 75.5% [4], 78.5% [5], and 88.4% [1].

We continue our work by designing REST-based web service to detect Indonesian language spam comments for Instagram using k-NN. This system is built on top of Amazon

This research is sponsored by LPPM Duta Wacana Christian University

Web Service (AWS) to provide a robust, resilience, and scalable system [14] as a foundation for our future work, which is a browser plugin.

We are using REST-based web services since it is simple, low usage on resources, and using HTTP protocol which is widely supported in many cloud service providers. We will conduct testing using assertion as described by Wenhui et al [15].

B. *Amazon Web Service*

Amazon Web Services (AWS) is a business subsidiary of Amazon Inc. providing cloud computing services that were started in 2006 and has become a market leader in cloud computing area. Recently, AWS was declared as a leader in "Magic Quadrant for Cloud Infrastructure as a Service" for 7 consecutive since 2010 [16].

AWS has a number of products that are divided into several categories: Compute, Storage, Database, Migration, Networking & Content delivery, Machine Learning, etc. AWS global architecture is depicted in Fig. 1.

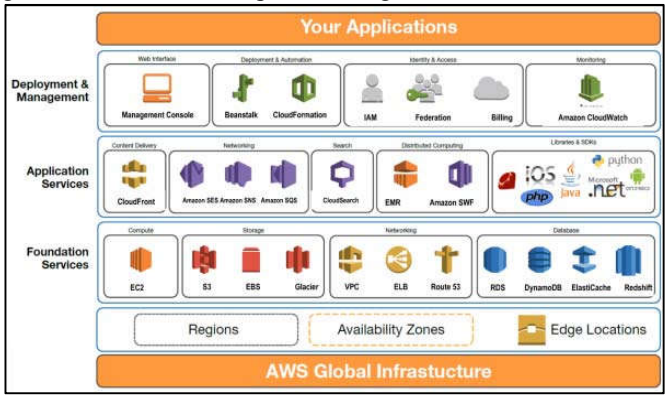


Fig. 1. AWS Architecture [17]

C. *Web Services (REST)*

Web service is a web-based service containing functions that represents the specific task. Web service uses one of the following technology:

1. Remote Procedure Call (RPC)
2. Simple Object Access Protocol (SOAP)
3. Representational State Transfer (REST)

REST Web Services is an architectural design for communication between applications through Internet using HTTP protocol [18]. Several HTTP operations used in REST are GET to fetch resources, PUT to update data, POST to create new resources, and DELETE to delete resources.

Since REST is running on top of the stateless protocol (HTTP), it can gain several advantages, especially related to cloud computing technology. Stateless components can be substituted easily in cloud computing since it does not require any data synchronization. Transactions or requests from clients are not required to be saved for future transactions. REST has several advantages against SOAP, which are better throughput and response time. REST is also more flexible compared to SOAP. [19].

D. *k-NN (k-Nearest Neighbor)*

K-Nearest Neighbor (k-NN) is a method that utilizes statistic principle to find the closest neighbor in the data. In supervised learning, k-NN is often called lazy learner because it does not learn anything from the data in the beginning, but it learns directly while doing classification process. K-NN works by finding some adjacent "k" data object or patterns based on the input and then choosing a class with the highest number of pattern in those k patterns. In short, k-NN using the voting principle to classify a pattern as in Fig. 2. The closest “k” pattern is decided by distance, similarity or dissimilarity, depending on its attributes.

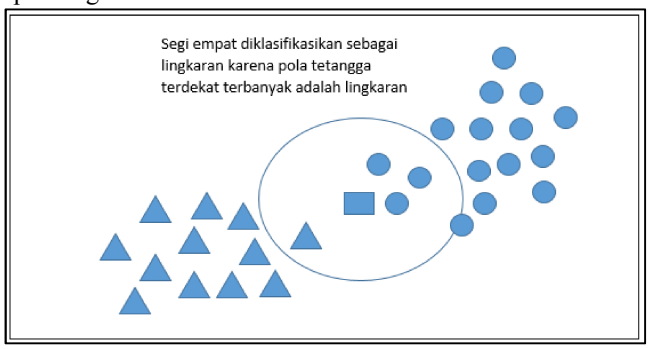


Fig. 2. K-NN Algorithm

K-NN has some advantages [20]:

1. Simple and easy to implement
2. K-NN works a locally by predicting k data until it matched with the locally clustered set of data

But it also has some disadvantages [21]:

1. Not efficient in terms of time because it always performs comparison operation without saving the information, thus not efficient for larger scale of data.
2. K values cannot be decided mathematically, thus further manual observation is still needed.
3. An even value of k also makes system cannot make a good decision. Lower k value will cause generalization, while higher k value will cause overfitting.

III. METHODOLOGY

We started our work by reviewing the AWS global architecture and look for the documentation about what we needed to build and deploy our applications into the AWS environment. Next, we started to design our architecture as in Fig. 3. We are using the compute and storage service mainly in this work, but we are adding more services in the future to ensure the reliability of the system under a production scenario. Our work is using a Java-based REST web services framework called Jersey. It is deployed on top of Apache Tomcat 8 as our main web server.

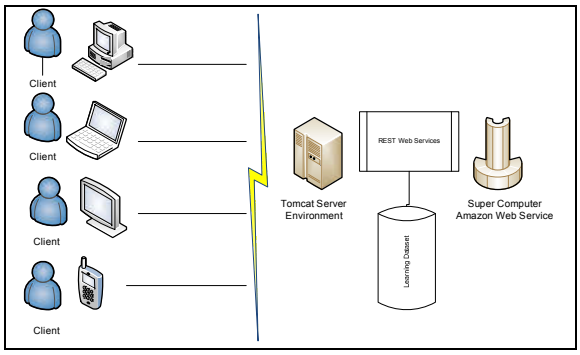


Fig. 3. Draft of Web Service Architecture

Next, we create the learning dataset by using our previous work. The format is in CSV file and it contains several attributes, such as actress ID, user, other attributes that are defined as features, and class (category). The records in CSV file are the weight per document according to their attributes. Table 1 and 2 shows the dataset and training data containing TF-IDF in form of CSV which will be uploaded into the AWS storage service (AWS S3).

TABLE 1. COMMENT DATASET IN CSV FORMAT

id_artist	user	comments	category
522969993	andreasnikoplak	butuh followers instagram atau likes instagram kamu sedikit yuk cek instagram unicorn apps stores melayani dengan ramah	Spam
522969993	nasardigunawan	Indonesia followers com udah tau tuh bisa nambahin followers	Spam
522969993	agvirasalami	mirip bapaknya cantik	Notpsam
…dst	…dst	…dst	…dst

TABLE 2. SYSTEM TRAINING DATA

Id	boikot	pacaran	pasif	payudara	…	Kategori
1	0.00505481	0	0	0.00505481	…	Spam
2	0.00607481	0.01705481	0	0	.	Nonspam
3	0.430001	0.00345	0	0	…	Nonspam

When designing the REST web service, we decided to use JSON format as the data-interchange format. We designed to build 5 functions:

- **classify**
  It is used to classify a document / new data in POST method. The input of this function is a document that the user wants to classify and a token.
- **version**
  It is used to show the web service version; implemented in GET method. This function does not require any input.
- **url**
  It is used to show the URL of the learning dataset currently being used; implemented in GET method. This function does not require any input.
- **file**
  It is used to retrieve an active dataset file currently being used in a text format; implemented in GET method. This function does not require any input.
- **dataset**
  It is used to replace dataset file number used as system training data; implemented in GET method. This function requires 1 input, which is dataset number, ranging from 1 – 8.

The last step is to implement the web service and deploy it to the AWS. We used Jersey framework (https://jersey.github.io/) while the k-NN is implemented using Kd-tree Java library (https://home.wlu.edu/~levys/software/kd/). We use the Eclipse equipped with AWS Toolkit plugin as our main IDE.

IV. RESULTS AND DISCUSSIONS

The result of this work is as follow:

1. **A working architecture deployed on top of AWS service**
   The final result of our architecture is depicted in Figure 4.

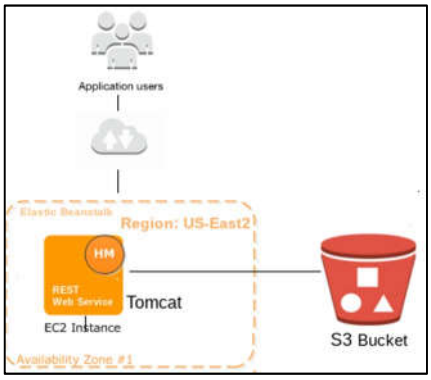


Fig. 4. Web Service Architecture

We deployed Elastic Beanstalk in the Ohio region (US-East2) and we used a Tomcat as our main server. All the dataset are kept in the S3 Bucket storage for durability and performance reason. There is only one instance at this point, but we are expecting to add more instances placed on multiple availability zones and region and setup a load balancer to distribute the load to several instances when the system is ready for production environment.

2. **Supervised Learning Dataset**
   We are using k-NN algorithm for supervised learning classification; thus, a training dataset is required. We have generated 8 training datasets using our own application that was built using PHP and

RapidMiner for this purpose. This enables users to select which datasets being used by using noDataset parameter in the request parameter.

Kd-tree Java library is selected for the k-NN implementation but we found that it has a limitation of handling a maximum of 1024 attributes in one dataset. This drawback forced us to limit every dataset to contain 1000 attributes at maximum. The details for each training datasets is described in Table 3.

TABLE 3. TRAINING DATASET

Code	Dataset	Description
1	PHP-unbalanced-non stem (16000 x 1000)	Number of data: 16000 Number of attributes: 1000 Generated manually by PHP-based system. This dataset contains unbalanced number of spam and non-spam data and they are not yet stemmed.
2	PHP-unbalanced-stem (16000 x 1000)	Number of data: 16000 Number of attributes: 1000 Generated manually by PHP-based system. This dataset contains unbalanced number of spam and non-spam data and they have been stemmed.
3	PHP-balanced-non stem (16000 x 1000)	Number of data: 16000 Number of attributes: 1000 Generated manually by PHP-based system. This dataset contains balanced number of spam and non-spam data and they are not yet stemmed.
4	PHP-balanced-stem (16000 x 1000)	Number of data: 16000 Number of attributes: 1000 Generated manually by PHP-based system. This dataset contains balanced number of spam and non-spam data and they have been stemmed.
5	Rapidminer-unbalanced-non stem (16000 x 1000)	Number of data: 16000 Number of attributes: 1000 Generated by RapidMiner. This dataset contains unbalanced number of spam and non-spam data and they are not yet stemmed.
6	Rapidminer-unbalanced-stem (16000 x 1000)	Number of data: 16000 Number of attributes: 1000 Generated by RapidMiner. This dataset contains unbalanced number of spam and non-spam data and they have been stemmed.
7	Rapidminer-balanced-non stem (16000 x 1000)	Number of data: 16000 Number of attributes: 1000 Generated by RapidMiner. This dataset contains balanced number of spam and non-spam data and they are not yet stemmed.
8	Rapidminer-balanced-stem (16000 x 1000)	Number of data: 16000 Number of attributes: 1000 Generated by RapidMiner. This dataset contains balanced number of spam and non-spam data and they have been stemmed.

3. Java Jersey REST-Based Web Service Implementation

The following are the endpoints for the functional web services:

1. version: http://tomcat.zdbgmwikga.us-east-2.elasticbeanstalk.com/service/version (GET)
2. no: http://tomcat.zdbgmwikga.us-east-2.elasticbeanstalk.com/service/no (GET)
3. url: http://tomcat.zdbgmwikga.us-east-2.elasticbeanstalk.com/service/url (GET)
4. file: http://tomcat.zdbgmwikga.us-east-2.elasticbeanstalk.com/service/file (GET)
5. dataset: http://tomcat.zdbgmwikga.us-east-2.elasticbeanstalk.com/service/dataset (GET)
6. classify: http://tomcat.zdbgmwikga.us-east-2.elasticbeanstalk.com/service/classify (POST)

All endpoints have been tested and the results can be seen in Fig. 5-10.

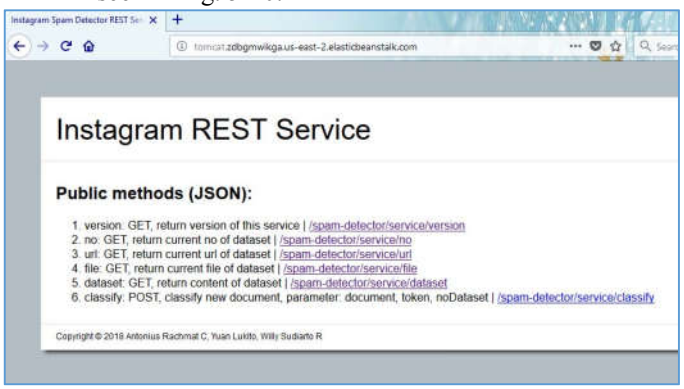


Fig. 5. Web Service Interface

The result of version endpoint using REST GET method is depicted in Fig. 6.

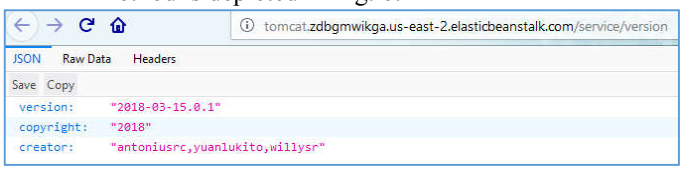


Figure 6. version() function result

The returned data is in form of JSON format as follows: `{"version":"2018-03-15.0.1","copyright":"2018","creator":"antoniusrc,yuanlukito,willysr"}`.
The result of noDataset endpoint using REST GET method is depicted in Fig. 7.

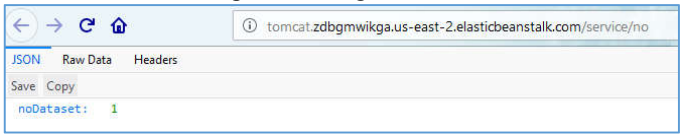


Figure 7. noDataset() function result

The returned data is in form of JSON format as follows: `{"noDataset":1}`.
The result of url endpoint using REST GET method is depicted in Fig. 8.

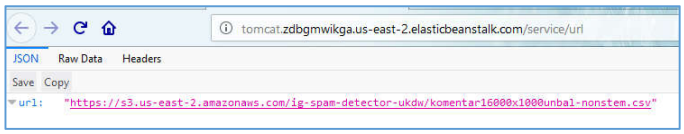


Fig. 8. URL() function result

The returned data is in form of JSON format as follows: `{"url":"https://s3.us-east-2.amazonaws.com/ig-spam-detector-ukdw/komentar16000x1000unbal-nonstem.csv"}`.
The result of getDataset using REST GET method is depicted in Fig. 9.

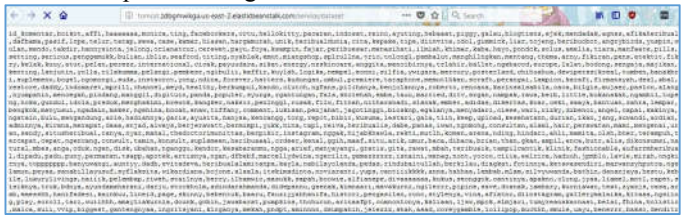


Fig. 9. getDataset() function result

The returned data from classify endpoint using REST POST method with comments, token, and noDataset parameters is depicted in Fig. 10.

TABLE 4. RESPONSE TIME

Test	Response Time (ms)					
	F1	F2	F3	F4	F5	F6
1	448	700	734	684	17000	744
2	330	624	622	628	10704	1672
3	332	484	1054	338	3726	640
4	694	780	636	690	3978	2554
5	624	652	620	1272	4172	802
6	600	628	1604	800	3680	620
7	322	598	1720	792	3412	804
8	602	800	486	334	3490	614
9	678	1628	646	642	3696	4986
10	328	668	786	702	4472	612
AVG	**495.8**	**756.2**	**890.8**	**688.2**	**5833**	**1404.8**
ST DEV	**158.47**	**319.26**	**433.68**	**262.08**	**4490.40**	**1407.63**

The overall average for response time is 1678.133 ms with deviation standard of 1178.6 ms.

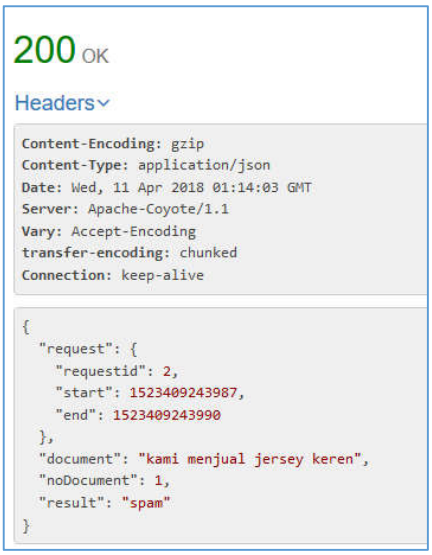


Fig. 10. classify function result

4. Web Services testing

After all web service endpoints are completed, we conducted performance test to see the response time in each service (F1 – F6). We are using RESTClient Firefox plugin (https://addons.mozilla.org/en-US/firefox/addon/restclient/) to conduct the testing and measure the results. The results of the response time tests are displayed in Table 4. Since AWS has many regions available, deploying on different regions might give a different result. Another possible solution to improve performance is to deploy AWS CloudFront, a global content delivery network that can deliver data with lower latency and higher transfer speed since requests will be redirected to closest edge location.

V. CONCLUSION

In this work, we have designed and developed a web service built using Java Jersey on top of Amazon Web Service (AWS). We tested the performance of this web service and the average response time is 1678.133 ms with deviation standard of 1178.6 ms.

We realized that further optimization is still required along with improvement to the number of attributes the system able to handle, improvement to kd-tree data structure, improvement of accuracy against a variety of provided datasets, and response time measurement when the environment is deployed in multiple regions or with CloudFront.

VI. ACKNOWLEDGMENT

This research has been funded by UKDW Research and Community Service (LPPM) unit under grant no. 072/D.01/LPPM/2018. The author also wished to thank Faculty of Information Technology Duta Wacana Christian University for their support.

VII. REFERENCES

[1] A. R. Chrismanto and Y. Lukito, "KLASIFIKASI KOMENTAR SPAM PADA INSTAGRAM BERBAHASA INDONESIA MENGGUNAKAN K-NN," in *Seminar Nasional Teknologi Informasi Kesehatan (SNATIK) 2017*, Yogyakarta, 2017.

[2] M. Eskelinen, "How to get rid of spam comments on Instagram," Mervi Eskelinen, 21 May 2013. [Online]. Available: http://merviemilia.com/blog/2013-05-how-to-get-rid-of-spam-comments-on-instagram. [Accessed 9 August 2017].

[3] D. Tamir, "How To Protect Yourself From Instaspam," 2015. [Online]. Available: http://readwrite.com/2015/04/15/instagram-spam-instaspam-how-to-avoid/. [Accessed 24 January 2018].

[4] A. Rachmat and Y. Lukito, "Deteksi Komentar Spam Bahasa Indonesia Pada Instagram Menggunakan Naive Bayes," *Ultimatics,* vol. 9, no. 1, 2017.

[5] A. R. Chrismanto and Y. Lukito, "Identifikasi Komentar Spam Pada Instagram," *Lontar Komputer,* vol. 8, no. 3, pp. 219-231, 2017.

[6] A. H. Wang, "Don't follow me: Spam detection in Twitter," in *Proceedings of the 2010 International Conference in Security and Cryptography (SECRYPT)*, Athens, Greece, 2010.

[7] A. H. Wang, "Detecting Spam Bots in Online Social Networking Sites: A Machine Learning Approach," in *IFIP Annual Conference on Data and Applications Security and Privacy*, Berlin, Heidelberg, 2010.

[8] M. McCord and M. Chuah, "Spam Detection on Twitter Using Traditional Classifiers," in *International Conference on Autonomic and Trusted Computing*, Banff, Canada, 2011.

[9] E. Tan, L. Guo, S. Chen, X. Zhang and Y. Zhao, "UNIK: unsupervised social network spam detection," in *CIKM '13 Proceedings of the 22nd ACM international conference on Information & Knowledge Management* , San Francisco, California, USA, 2013.

[10] J. Xin, J. Luo, C. X. Lin and J. Han, "A Data Mining-based Spam Detection System for Social Media Networks," in *Proceedings of the VLDB Endowment*, Seattle, Washington, 2011.

[11] X. Zhang, S. Zhu and W. Liang, "Detecting Spam and Promoting Campaigns in the Twitter Social Network," in *2012 IEEE 12th International Conference on Data Mining (ICDM)*, Brussels, Belgium, 2012.

[12] K. Hines, "How to Identify and Control Blog Comment Spam," 2012. [Online]. Available: https://blog.kissmetrics.com/control-blog-comment-spam/.

[13] G. Mishne, C. D and L. R., "Blocking Blog Spam with Language Model Disagreement," in *Proceedings Of The First International Workshop On Adversarial Information Retrieval On The Web (Airweb)*, 2005.

[14] C. Radford, "Challenges and solutions protecting data within Amazon Web Services," *Network Security,* vol. 2014, no. 6, pp. 5-8, 2014.

[15] H. Wenhui, H. Yu, L. Xueyang and X. Chen, "Study on REST API Test Model Supporting Web Service Integration," in *2017 IEEE 3rd International Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, Beijing, China, 2017.

[16] L. Leong, R. Bala, C. Lowery and D. Smith, "Gatner," 15 June 2017. [Online]. Available: https://www.gartner.com/doc/reprints?id=1-2G2O5FC&ct=150519&st=sb&aliId=1221201968.

[17] J. Townsend, "Introduction to Amazon Web Services (AWS)," 20 Juy 2015. [Online]. Available: http://vmtoday.com/2013/07/introduction-to-amazon-web-services-aws/.

[18] R. Fielding, Architectural Styles and the Design of Network-based Software, California: University of California, 2000.

[19] S. Kumari and S. K. Rath, "Performance comparison of SOAP and REST based Web Services for Enterprise Application Integration," Kochi, India, 2015.

[20] Suyanto, Data Mining untuk Klasifikasi dan Klasterisasi Data, Bandung: Informatika, 2017.

[21] S. Jiang, G. Pang, M. Wu and L. Kuang, "An improved K-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications,* vol. 39, no. 1, pp. 1503-1509, 2012.

Parking Space Detection Using Quaternionic Local Ranking Binary Pattern

Markus Antoni Suwignyo, Iwan Setyawan, Banu Wirawan Yohanes ‡
Faculty of Electronic and Computer Engineering
Satya Wacana Christian University
Salatiga, Indonesia
Email: 612013041@student.uksw.edu, iwan.setyawan@ieee.org, ‡banu.yohanes@staff.uksw.edu

***Abstract*—In the digital age, some places have been installed with a surveillance camera, especially in a parking lot. In order to make use of that camera, a parking space detection is developed to help users find empty spaces. With so many paper that also studied this kind of system, we want to explore another feature extraction that could be used in this system. One particular feature extraction algorithm that we selected is quaternionic local ranking binary pattern (QLRBP). We also used local binary pattern (LBP) to compare it's performance. The result shows that QLRBP is able to achieve the accuracy score above 99%, but with expense of large feature vector size.**

***Keywords*—LBP, QLRBP, parking detection, image processing**

I. INTRODUCTION

Finding an empty space sometimes become our daily life problems. This is because in the city, parking slot is limited and hard to find. But, in the digital age, some places have been installed with a surveillance camera, especially in a parking lot. In order to make use of that camera, a parking space detection is developed to help users find empty spaces.

In general, parking space detection has 3 parts. The first one is acquisition, which responsible to capture an image from the real world. The second is feature extraction, which responsible to extract a meaningful information from the image. The last one is classifier, which responsible to classify the image into one of several classes.

Several authors have been proposed such system and their result is surprisingly good [1]–[6]. For example, Qi Wu et al [1] use Gaussian Ground Color Model as feature extraction and Support Vector Machine as classifier. Another example is Almeida et al [6] that used Textural Descriptor as it's feature extraction and Support Vector Machine as it's classifier. Even so, we want to try another possible feature extraction that serve as alternative towards previous work.

In this paper, QLRBP is used as feature extraction for parking space detection. QLRBP is an extension of LBP that use color image instead of using grayscale image. LBP itself is widely used as feature extraction in face recognition. It's capability to describe the texture of image has able to create a discriminative feature better than PCA [7]. In II-B, it has been shown the advantage of QLRBP. We believe that QLRBP has more discriminative feature than LBP, therefore could result a better prediction.

We also used k-nearest neighbor (knn) and support vector machine (svm) as our classifier algorithm. To validate the result, 10-fold validation method is used.

We want to show that QLRBP can be used in parking space detection as feature extraction. Also, we want to prove that QLRBP is much superior than LBP.

II. FEATURE EXTRACTION

In this section, we will describe briefly about feature extraction that we will use in this experiment.

A. Local Binary Pattern

LBP was first proposed by Ojala et al and since then LBP has been used widely in *face recognition* [8]. LBP is also called a textural descriptor because LBP could capture the texture from the image. The advantage using LBP is resistant to illumination variation. The basic LBP operator compares the eight neighbouring pixel grayscale values to the centre one, assigning a 0 or 1 to each neighbour depending on whether they are less than or greater than the centre value. Usually, these neighbouring pixel are defined inside 3×3 block S_m with x_m as the center pixel. These binary values are then concatenated in predetermined order and converted to a single decimal number that represents the "type" of local image structure [9]. Mathematically,

$$\text{LBP}_{x_m} = \sum_{n=0}^{|S_m|} h(x_n - x_m), \quad h(t) = \begin{cases} 1 & t \geq 0 \\ 0 & t < 0 \end{cases} \tag{1}$$

where $|S_m|$ is the number of elements in S_m. Fig 1 shown the example on how LBP coding works.

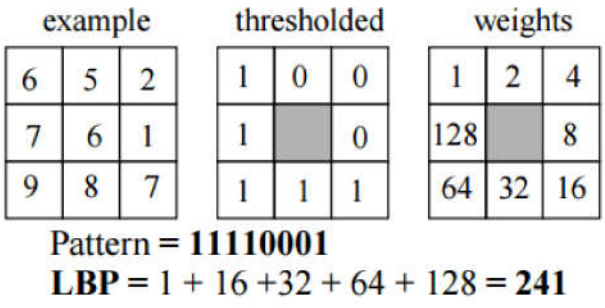


Fig. 1. Example on how LBP coding works

This operator then will transform the image into another image with a different value. The resulting image itself desribe more about texture of the image and have corner detection property. Below is a sample image after LBP operation from image 8.

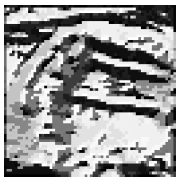


Fig. 2. The resulting image after LBP operation from image 8

We hypothesize that LBP could make histogram information more discriminative towards vacant and vacuum parking space.

B. Quaternionic Local Ranking Binary Pattern

QLRBP is yet another extension method from LBP. Different from traditional descriptors that are extracted from each colour channel separately or from vector representation, QLRBP works on quaternionic representation (QR) of the colour image that encodes a colour pixel using a quaternion. QLRBP is able to handle all colour channels directly in the quaternionic domain and include their relation simultaneously. Fig 3 shows the framework of QLRBP in simplified manner.

The following is the explanation of the framework itself.

1) From color image, we encode each value with QR. For example, RGB value of `(200, 200, 200)` will be encoded into $0 + 0.78i + 0.78j + 0.78k$.
2) Choose $\dot{p}_1$, $\dot{p}_2$, and $\dot{p}_3$ as reference quaternion. To design reference quaternions, we could:
   a) Encode a color image using vector representation, and apply k-means algorithm to find clusters as the reference quaternions.
   b) Develop hand-crafted reference quaternions with only one complex operator, where two elements are set to zero . In other word, $\dot{p}_1 = i$, $\dot{p}_2 = j$, and $\dot{p}_3 = k$.
   c) Develop hand-crafted reference quaternions with one dominant complex operator, where two elements are set to small random number $\epsilon = [0, 0.1]$.
3) Perform CTQ to all elements by $\dot{p}_i$, then calculate the phase image of the Clifford translated image CTQ. Mathematically, CTQ is defined as follows:

$$\text{CTQ}_l(\dot{p}, \dot{q}) = \dot{p}\dot{q} \tag{2}$$

Phase image is obtained with this definition:

$$\theta = -\arctan\left(\frac{\sqrt{b^2 + c^2 + d^2}}{a^2}\right) \tag{3}$$

4) To rank two quaternion $\dot{q}_m$ and $\dot{q}_n$ we use this function:

$$R_{\text{QLRBP}}(\dot{q}_m, \dot{q}_n) = \theta_{\text{CTQ}(\dot{q}_n, \dot{p}_t)} - \theta_{\text{CTQ}(\dot{q}_m, \dot{p}_t)} \tag{4}$$

where $t \in 1, 2, 3$.
5) Based on QR and the ranking function in Eq. 4, we can derive the definition of the QLRBP coding. Considering an image pixel $\dot{q}_m$ centered in a 3×3 block S_m and a specified reerence quaternion $\dot{p}_1$, the QLRBP coding is calculated by:

$$\text{QLRBP}_{\dot{q}_m} = \sum_{n=0}^{|S_m|-1} h(R_{\text{QLRBP}}(\dot{q}_m, \dot{q}_n))2^n \tag{5}$$

where $|S_m|$ is the number of elements in S_m, $h(\cdot)$ and $R_{\text{QLRBP}}(\dot{q}_m, \dot{q}_n)$ are defined in Eqs. 1 and 4.
6) Similar to LBP, the histogram of the QLRBP coding image is used as features for classification. In order to obtain more robust features against variations of illumination, view angle, pose, and occlusion, the histogram is extracted from overlapping cells like other dense local. The centers of all cells are uniformly distributed within the image. A normalization is then performed to the histogram generates the QLRBP feature representation of the original color image.

From [10], QLRBP has the following advantages.

1) QLRBP is based on QR of the color image. Encoding all color channels together, QR offers a flexible way to handle them at the same time. Several quaternionic operators can be used to analyze the color image in the quaternion space. Therefore, the relations between different color channels are considered and rich information can be extracted from the original color image.
2) CTQ is used in QLRBP. Each portion of the CTQ result has its specific meaning in the original color space, i.e. the real part describes the similarity of all color channels of two pixels, while the imaginary part consists of the differences between two different color channels.
3) Compared to local color vector binary pattern (LCVBP) [11]

Figure 4, 5, and 6 are sample images after QLRBP operation from image 8 with various techniques method for choosing reference quaternion.

III. Experiments

A. Image Data

Image data that we used are taken from Federal University of Parana in Brazil [12]. Those images has 3 different parking space, that is UFPR04 (28 parking slot), UFPR05 (45 parking slot), and PUCPR (100 parking slot). In this experiment we used spesifically only UFPR04. There's a 3791 images taken (or in total 106,408 parking slot), but for convenience, we only consider 6,000 parking slot images that has been transformed first (with perspective transform into 64×64 image). We choose the images with random and already balanced the classes between vacant and vacuum parking slot.

Since we use 10-fold validation, we divide those 6000 image into 10 subsamples. In this validation, one experiment will repeated 10 times. Each iteration use 1 subsamples for



Fig. 3. Framework for QLRBP



Fig. 4. The resulting image after QLRBP operation from image 8 where reference quaternions determined with k-mean algorithm



Fig. 5. The resulting image after QLRBP operation from image 8 where reference quaternions determined hand-crafted with only one complex operator

training sample and the rest of it for testing data. We guarantee that all iteration has different training data.

B. Performance Evaluation

In this work, we only evalute the performance of both extraction feature by using accuracy score. The definition of accuracy score is as follows

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{6}$$

where TP, FP, FN, and TN are True Positive, False Positive, False Negative, and True Negative, respectively.

C. Experiment 1

1) Experimental Setup: Lan et al. works shown QLRBP and applied those algorithm in *person reidentification*. The experimental results demonstrated the effectiveness of the proposed QLRBP. In our works, we apply QLRBP into parking



Fig. 6. The resulting image after QLRBP operation from image 8 where reference quaternions determined hand-crafted with one dominant complex operator



Fig. 7. Example of parking lot image



Fig. 8. Example of parking space image example

space detection system. But first, we need to know the best parameter QLRBP for this kind of system. For this experiment, we tried various parameter, such as overlapping-cell size and technique for choosing reference quaternions.

First, we build our system with Python and some useful package like OpenCV (for image processing) and scikit-learn (for classifier). After that, we feed those system with the image database and get the accuracy score. In this experiment we build one system (QLRBP + knn) with some defined parameter that we would like to observe. There's 6 variation parameter that we used.

- Technique: k-mean algorithm, overlapping-cell size: 32×32
- Technique: hand-crafted; only one complex operator, overlapping-cell size 32×32
- Technique: hand-crafted; one dominant complex operator, overlapping-cell size: 32×32
- Technique: hand-crafted; one dominant complex operator, overlapping-cell size: 16×16
- Technique: hand-crafted; one dominant complex operator, overlapping-cell size: 16×16

2) Experimental Result: By using knn ($k = 3$) as classifier, the best result were achieved using 32×32 overlapping-cell size and hand-crafted reference quaternions with one dominant complex operator and small random number on others. Table I and II shown the experiment results.

TABLE I. ACCURACY SCORE EACH SYSTEM WITH VARYING TECHNIQUE FOR CHOOSING REFERENCE QUATERNIONS

Technique for choosing reference quaternions	Accuracy score
k-mean algorithm	97.367%
hand-crafted; only one complex operator	98.85%
hand-carfted; one dominant complex operator	98.933%

TABLE II. ACCURACY SCORE EACH SYSTEM WITH VARYING OVERLAPPING-CELL SIZE

Overlapping-cell size	Accuracy score
32×32	98.933%
16×16	98.733%
8×8	95.05%

It turns out, by making overlapping-cell size smaller, it only makes the accuracy score worse. We guessed that the smaller the size is, the more information the we got from QLRBP. Of course more information doesn't mean the classifier would achieve high accuracy since maybe there's some redundant information in there. This could cause higher computational cost on training phase and no guarantee that classifier could perform better.

Another thing that we want to point out is that hand-crafted technique performs better than k-mean algorithm. We believe that choosing reference quaternion dynamically might result a redundant information. This is also a possible case when the reference quaternion that where picked are almost the same. By choosing the reference quaternion manually, we could avoid those problem. The result also shown that the 3rd technique is slightly improving the 2nd technique. Lan et all stated that although two value on other complex operators is set to a small value, it helps to preserve interactions between color channels. [10]

To assert our previous paragraph, take a closer look on figure 4, 5, and 6. If we look carefully on figure 4, we found that 2 image is almost identical. This is often happen when we choose reference quaternion dynamically with k-mean algorithm. We could assume that two reference quaternion is simply almost identical. Meanwhile, figure 5 and 6 has a different image output. This is happen because we choose reference quaternion statically with hand-crafted value. Of course we could assume the reference quaternion on each image is different.

D. Experiment 2

1) Experimental Setup: After we know the best QLRBP's parameter, the next experiment is to compare perfomance of LBP and QLRBP. We must address that LBP algorithm that we used in this work is the basic one. The generalized version of LBP allows us to set P, another notation for $|S_m|$, and R, block S_m radius. Ojala et al. also create another flavor of LBP like LBP uniform pattern and LBP rotation invariant that is beyond on our work. [8]

Like previous experiment we build our system with Python and some useful package like OpenCV (for image processing) and scikit-learn (for classifier). After that, we feed those system with the image database and get the accuracy score. In this experiment we build four system, LBP + knn, LBP + svm, QLRBP + knn, and QLRBP + svm.

For knn, we tried various k with $0 < k < 50$ (only odd number). For svm, we tried various C with $10^{-3} \leq C \leq 10^{7}$ and y with $10^{-5} < \gamma < 10^{3}$ (logarithmic grid). But in this paper, we only present the best one.

2) Experimental Result: The classifier that we used in this work are knn and svm. Various parameters were tried for both knn and svm. For knn, we tried various k and the best result were achieved using $k = 3$. For svm (with *radial basis function* kernel), we also tried various C and γ. The best result were achieved using $C = 10$ and $\gamma = 10^{-5}$. The end result of this experiment are shown on Table III

TABLE III. ACCURACY SCORE OF ALL CLASSIFIER USED IN THIS WORK

Feature Extraction	Accuracy Score	
	with knn	with svm
LBP	98.1%	98.967%
QLRBP	99.183%	99.35%

From the table above, we can see the superior of QLRBP. Both knn and svm can achieve 99% accuracy, whilst LBP can't. But, on the side note, we must notice that QLRBP feature vector size is bigger than LBP. This is because QLRBP use overlapping-cell and 3 reference quaternion. We know that the bigger feature vector size is the longer computational process on classifier would be.

IV. CONCLUSION

In this paper, we shown LBP and QLRBP as feature extraction on parking space detection. The result has shown that QLRBP has achieved higher accuracy score than LBP. But to achieve such high accuracy, QLRBP used larger feature vector size than LBP, that could burden the computational cost on classifier.

ACKNOWLEDGMENT

K-mean algorithm used here is supported by UKSW internal research. We would also like to thank UKSW for their support for this publication.

BIBLIOGRAPHY

[1] Q. Wu, C. Huang, S.-y. Wang, W.-c. Chiu, and T. Chen, "Robust parking space detection considering inter-space correlation," in *Multimedia and Expo, 2007 IEEE International Conference on*, IEEE, 2007, pp. 659–662.

[2] R. J. L. Sastre, P. G. Jimenez, F. J. Acevedo, and S. M. Bascon, "Computer algebra algorithms applied to computer vision in a parking management system," in *Industrial Electronics, 2007. ISIE 2007. IEEE International Symposium on*, IEEE, 2007, pp. 1675–1680.

[3] D. Bong, K. Ting, and K. Lai, "Integrated approach in the design of car park occupancy information system (coins).," *IAENG International Journal of Computer Science*, vol. 35, no. 1, 2008.

[4] C.-C. Huang, S.-J. Wang, Y.-J. Change, and T. Chen, "A bayesian hierarchical detection framework for parking space detection," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, IEEE, 2008, pp. 2097–2100.

[5] C.-C. Huang and S.-J. Wang, "A hierarchical bayesian generation framework for vacant parking space detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 12, pp. 1770–1785, 2010.

[6] P. Almeida, L. S. Oliveira, E. Silva, A. Britto, and A. Koerich, "Parking space detection using textural descriptors," in *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*, IEEE, 2013, pp. 3603–3608.

[7] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[8] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[9] S. J. Prince, *Computer vision: models, learning, and inference*. Cambridge University Press, 2012.

[10] R. Lan, Y. Zhou, and Y. Y. Tang, "Quaternionic local ranking binary pattern: A local descriptor of color images," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 566–579, 2016.

[11] S. H. Lee, J. Y. Choi, Y. M. Ro, and K. N. Plataniotis, "Local color vector binary patterns from multichannel face images for face recognition," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2347–2353, 2012.

[12] P. R. De Almeida, L. S. Oliveira, A. S. Britto, E. J. Silva, and A. L. Koerich, "Pklot–a robust dataset for parking lot classification," *Expert Systems with Applications*, vol. 42, no. 11, pp. 4937–4949, 2015.

A Comparison Study of Search Strategy on Collecting Twitter Data for Drug Adverse Reaction

Arif Dwi Laksito[1)], Kusrini*[2)], Heri Sismoro[3)], Fita Rahmawati[4)], Mochammad Yusa[5)]
[1,2,3,5)]Universitas AMIKOM Yogyakarta
Jl. Ringroad Utara Condong Catur Depok Sleman Yogyakarta Indonesia
[4)]Universitas Gadjah Mada Yogyakarta
Bulaksumur, Caturtunggal, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta Indonesia
Email: [1)]arif.laksito@amikom.ac.id, [2)]kusrini@amikom.ac.id, [3)]herisismoro@amikom.ac.id [4)]rahmawati_f@ugm.ac.id
[5)]mochammad.yusa@amikom.ac.id

***Abstract*- A changes in the effects of drugs can be occured because the patient consumes two (or more) drugs or consumes them with a specific food or beverage resulting an unexpected adverse reaction. The abundance of data in social media can become the based of ingredients that is very valuable in constructing drug interaction analysis knowledge for pharmaceutical experts to produce solution for this problem. The data in this study comes from social media Twitter via its standard search API and be collected using software tools "R" via its Twitter library. We observed two scenario strategies in collecting data in order to find the best strategy between them that can retrieved largest number of result. Scenario 1 was based on a combination of two or more keywords, while the Scenario 2 uses only single keyword search mechanism followed by a filtering process. A list of drugs and food interaction related keyword was first declared before the process. There are 20 keywords to be tested, divided into 5 cases and categorized into 4 attributes for each case: Event (K), Drug (O), Effect (E) and Food Beverage (M). The experiment showed that Scenario 2 with single keyword and filtering strategy is the best keyword finding mechanism that collected as many as 1323 data, while the Scenario 1 that using keyword combination only collected 438 data.**

Keywords: Drugs interaction, twitter API, tools "R", Collecting data, Knowledge

I. INTRODUCTION

Drug interactions are the interaction among drugs that occur when a patient receives two or more drugs at the same time causing a change in the effect of consumed drugs. Drug-drug interactions can result in the occurrence of unexpected adverse events that can be fatal or even leads to death. In the treatment practices, there are many cases of treated patient in the hospital caused by the interaction of drugs consumed by patients, such as the case of hypoglycemia that occurs after the use of gliburid and co-trimoxazol simultaneously. Another example is the occurrence of toxicity of digoxin when used simultanously with clarithromycin. In addition to drug-drug interactions, it is also possible the interaction occurs between drugs to certain foods.

Drug interactions are the most common problem in drug treatment. The incidence of drug interactions increases following the amount of drug consumed. Such case occurs mostly to elderly patients who consume several drugs at once due to the illness that often followed by complications. The study of Rahmawati et al showed that from 100 inpatients, 65% of them got the drug interaction problems with a total number of drug-drug interactions 204[1].

Obtaining the knowledge about drug interactions can be done from many resources and using many strategies. Generally the knowledge is compiled from pharmacy experts by conducting a series of experiments whose results are then arranged into text book of drug interactions reference.

The rise of online social media generates a huge pile of data. The huge amount of data stacks opens an opportunity to do text analysis based on the user generated contents. Social media has been used as a source of data for many studies in areas such as in mental health that are used to process emotional content on a population in real time with interactive visualization[2] or in social fields such as to detect people's reaction to a growing news or issues[3]. Social media can also be used to provide recommendations to speed up decision making of tourists destinations[4], and as a recommendation for social media users in selecting and finding content that is appropriate and useful for users[5]. Stieglitz and Dang-Xuan analyzed the development of contents in several online medias to determine the context of politic organizations[9]. In that research, the data crawling processs from microbloging become the target to make framework analysis for political communication. One of the microblogging as the source of data collection is twitter. The collection of text data containing political elements is based on the names of politicians, specific keywords, or simply a random searches. However, the amount of data collected through the microblogging is not explained in this study. Grujic, Bogdanovic-Dinic, and Stoimenov (2014) extract text data from specific facebook pages on governmental area[10]. In this study, data collection is done based on the official facebook pages from various ministry institutions to make a recommendation about the best time to conduct socialization campaign related to a governmental field. Graph API technologies provided by facebook and twitter are slightly different. Twitter API provides a search function based on specific keywords. Liang et.al (2016) collected data using the twiiter API based on a specific user

list named CrowdAPS[11]. The study is limited only by relying on the twitter user list and still weak in handling data collection using certain keywords. Ferrara et.all (2014) applies the concept of web extraction to find text data based on user correlations and profiles on social media one of which is twitter[12]. In this study the effectiveness of the amount of data obtained has not been explained. Chang, Saravia, and Chen (2016) introduced Subconscious Crowdsourcing to collect mental disorder information on online media [13]. The data collection mechanism is based on three different attributes: random-sampled users, patients and experts by combining the value data of the three attributes manually.

Studies conducted by Stieglitz and Dang-Xuan [9], Grujic, Bogdanovic-Dinic, and Stoimenov [10], and Liang et.al[11], collect data based on the actor (actor-based) for later processing. In the search of drug to drug interaction data, it is very difficult to find social media user profiles that make posts related to drug and food interactions so that data collection using actor-based mechanisms is deemed ineffective. Ferrara et.all[12] introduced a mechanism of data collection on online social media using specific keyword techniques. In this case, for data collection of drug and food interaction is considered suitable to be implemented. The focus of the study is to calculate the effectiveness of data collection based on specific keywords using the annotation mechanism adopted from Chang, Saravia, and Chen [13]. The challenge in collecting data from social media is the limited amount of data provided by vendors [8].

Twitter is one of today's popular social networking media. Currently the number of tweets per second reaches 8000 data with the number of existing users reaching 330 million users[6]. As part of the business model of social media networking companies, they created an Application Programming Interface (API) to provide interaction with third parties. This allows researchers to collect data from social media to be retrieved, stored, manipulated for further analysis[7].

Currently Twitter provides several types of API for data search that is premium search, enterprise search and standard search. Unlike the first and second types, the standard search can be used for free. This research will study and select the method for data collecting from social media twitter in accordance with the need of knowledge of drug interaction. Researchers use standard search where the data obtained is ranged for the last 7 days and in the form of data sampling. The challenge in this study is how to create a method to maximize the appropriate search using standard search APIs on twitter[8].

The focus of this research will be to discuss the effectiveness of annotation mechanisms in order to obtain a high number of relevant data by searching data based on specific keywords. The study result will be useful for the future research such as construction of drugs interaction basic knowledge for drug interaction analysis.

II. RESEARCH METHOD

This research is an experimental research as pictured in Figured 1. Every data collection process of the scenario utilize the Twitter library of tools "R" version 3.4.3 to inquire the data from Twitter by its search API. The research begins by declaring a list of keywords related to drug and food interactions. Keywords are categorized by incident attribute (K), Drug (O), Effect (E), and Type of Food or Beverage (M) (Figure 1). Then the first test scenario is started by combining keywords based on attributes. The term "combine" is referred as "AND" operator in data retrieval. The combination made consists of a combination of two attributes (K-O, C-E, K-M, O-E, O-M, E-M), three attributes (K-O-E, K-E-F, O-E-F, K-O-M), and the four attributes (K-O-E-M). The tests were conducted on 5 cases related to drug and food interaction keywords (Table 1). After data collection step is completed, the next step is to calculate the total amount of data collected based on keyword combinations of each attribute. The result data will be the comparison to the amount of data obtained in the second scenario.

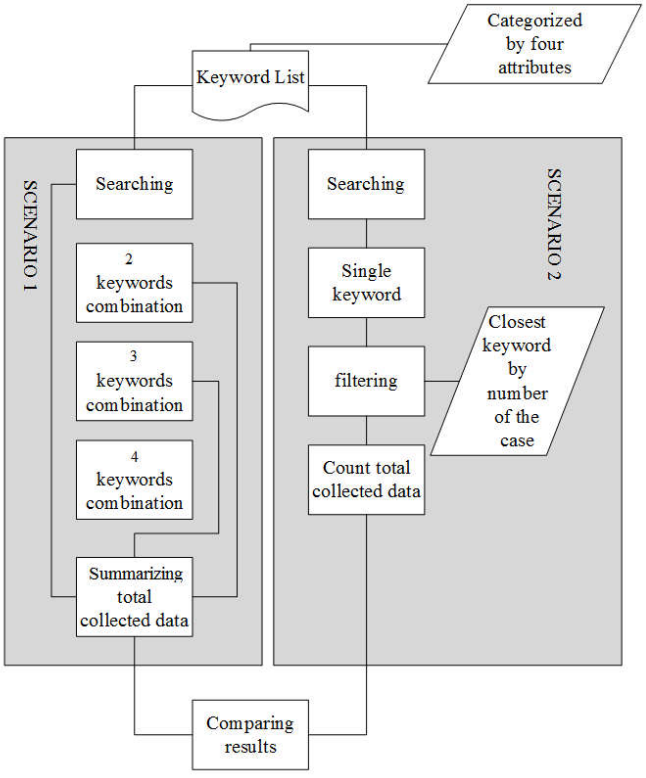


Figure 1. Research methodology

The second test scenario begins by collecting data based on single keyword from a list of drug and food interaction keywords. In this study we limit the amount of data collected for searches based on a maximum of 10,000 keywords. Once each keyword-related data is collected, the data will be filtered based on the keywords of each case. Filtering results will be calculated per single keyword. Then each case will be calculated relating how much data is collected and relevant or

contains two or more combinations of keywords. The result of scenario 2 will be compared with the result of previous scenario 1 to determine which one is more effective in terms of the sum of document that can be retrieved.

III. Result and Discussion

The focus of this research is to utilize the API from Online Social Media Twitter to obtain or collect data related to drug and food interactions. The research tests using the keyword that has been categorized by pharmacyst with number of case is 5 and each f the case containing 4 attributes that has been specified based on event of

The first stage is to determine the list of words based on event (K), drug (O), effects (E), and food or beverage (M). The list of predefined words can be seen in table 1. The keyword used is in Bahasa Indonesia as the implementation to be expected for Indonesia country.

Table 1. Keywords list

Case	Event (K)	Drug (O)	Effect (E)	Food (M)
1	*Keluhan*	*Amlodipin*	*Muntah*	*Sprite*
2	*Obat*	*Kapsul*	*Ngantuk*	*Semangka*
3	*Setelah Minum obat*	*Puyer*	*Muntah*	*Buah*
4	*Obat tidur*	*Serbuk*	*Gatal*	*Jus*
5	*Obat batuk pilek*	*CTM*	*Pusing*	*Es krim*

Table 1 represents a list of keywords that are then used for data collection from twitter. The approach used is explorative annotation method based on the specific keyword used in the research [13]. The data is collected by utilizing the API provided by the online social media vendor twitter. The data collection is based on certain keywords. The data collection mechanism in this study uses R language using library from Jeff Gentry [14].

The "twitteR" library from Jeff Gentry allows a data searching in twitter using parameter of text, data number, and language. The text parameter using the combination of 2 to 4 predefined keyword attributes.

```
> searchTwitter("Keluhan+Amlodiplin", n=10000, lang="id")
> searchTwitter("Keluhan Amlodiplin", n=10000, lang="id")
```

Figure 2. Searching function in twitteR library of "R"

The parameter of keywords attribute combination can be separated by using '+' or space character as seen in Figure 2. The result of those inquiries return all tweets data that containing "Keluhan" and "Amlodiplin" in them.

The first explorative approach used is an explorative approach based on word combinations of the food-interaction attributes (table 1). The approach is a combination of two or more values. Then used data collection. The first approach is the method of collection used is to combining based on keyword clusters. The combination is a combination of keywords at least two or more value combinations. For example the combination for case one is the combination of the word "keluhan" in the value of the event attribute with the keyword "amlodipin". Then the word combination will be used for collecting data. The amount of data obtained from the exploration method of combining two or more values can be seen in table 2.

Table 2. The result of data crawling collection with combination of 2 values

Case	K-O	K-E	K-M	O-E	O-M	E-M	SUM
1	4	3	0	0	0	15	**22**
2	150	146	3	0	0	0	**299**
3	0	0	0	0	15	16	**31**
4	0	0	15	63	0	0	**78**
5	3	0	1	0	3	0	**7**
TOTAL	**157**	**149**	**19**	**63**	**18**	**31**	**437**

Table 2 represents the data collected from Social Media Online twitter using two keywords. Keyword combinations are represented by case and keyword combinations for crawling. K represents the column or attribute of events (Table 1), O represents Drug, E represents Effect, and M is the type of food or beverage. The results of data retrieval experiments using a combination of the value of each attribute. The combination of keyword case 1 between data value of adn onbat events (K-E) collected relevant data for 4 tweets. Total data collected in case 1 were 22 data, case 2 was 299 data, case 3 were 31 data, case 4 was 78, and case 5 were 7 data. The total data collected using a combination of two keywords is 437 data.

Table 3. The result of data crawling collection with combination of 3 values

Case	K-O-E	K-E-M	O-E-M	K-O-M	SUM
1	0	0	0	0	0
2	1	0	0	0	1
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
TOTAL	**1**	**0**	**0**	**0**	**1**

Table 3 is the result of data crawling using a combination of 3 attribute values. The result of data collected as a whole is only one data. Where the combination of 3 values from case 1,3,45 returned no data. The only data collected was obtained from the combination of Keywords-Drug-Effects (K-O-E) keywords in case 2.

Meanwhile, as predicted, the combination of 4 attribute values (K-O-E-M), return no data collected from the five cases. As shown in Table 4, from the Scenario 1 the total data collected using combination 2 to 4 values is 438 data.

Table 4. The result of Scenario 1

Combination	SUM
2 values	437
3 values	1
4 values	0
TOTAL	**438**

The second approach used in Scenario 2 is to collect words that are general by case and then filter based on the four based on the four keywords of each attribute such as events, drugs, effects, type of food or beverage. Activity stages begin by collecting all data by keyword per case. Then each collected data will be converted to the database. After the data obtained per keyword, the next step is the merger of all data that has been crawled and sorted based on the key words per attribute of data that is crawled from online social media twitter. The sum of search results for each word can be seen in Table 5.

The total number of data collected is 93.976 data. In this study, the maximum amount of data collected by keyword is only 10.000 data. The limit is subject to standard version of twitter API that has maximum rate limit of 180 per request [15] in order to keep the respon time quite acceptable.

The next step is the process of filtering based on crossed cases. For example of cross filtering, if the filter is on the keyword *keluhan*, then the consideration will be on keywords *amlodipin, muntah,* and *sprite*. Also if the filtering on the keyword *sprite* then the consideration is on the keywords of *keluhan, amlodipin*, and *muntah*. In case 1 ,keyword *keluhan*, *muntah*, *amlodipin* and *sprite* is the keyword being considered to be filtered and searched for words that contain the nearest keyword. The same treatment is applied to case 2,3,4, and 5. The results and words that are considered common to the filtering process can be seen in Table 6.

After going through the filtering process of 10.000 data that contain word *keluhan*, none of the data that contain the word *Amlodipin, Muntah,* or *Sprite*. Whereas from 10.000 data containing word *muntah* only 2 data that have relationship with word *amlodipin, keluhan*, or *sprite*. From the 797 data that contained the word sprite, there is only 3 data contained a combination of sprites, amlodipin, keluhan and muntah. Amlodipin keyword do not display data that has relations with those three words. Total data collected after filtering process in case 1 is 5 data. The filtering process case 1 to 5 is treated with the same schema. The number of data collected from the second case is 852 of the total data as much as 27.100, the third case collected 30 data from a total of 20.316 data, the fourth case collected 420 data from total 18.952 data, and the fifth case of 16 data from total 16.808 data. So the total data collected is 1.323 data. If this data is compared to the word combination scheme performed on the first experiment the data collection results done using the second scheme get better results. In the first scheme collection using the total amount of data collected is as many as 438 data while the filtering experiment scheme, the data collected as much as 1.332 data. Therefore, with the limitations of API policies provided by standard twitter vendors that data collection is only done based on data 7 days back, it was concluded that the method of using the word combined filtering process will have a better value of effectiveness compared to the search scheme based on the number of keyword combinations.

Table 5. The Sum of Crawled Data Per Keyword

Classification	**Case 1**		**Case 2**		**Case 3**		**Case 4**		**Case 5**	
	Keyword	**Sum**	**Keyword**	**Sum**	**Keyword**	**Sum**	**Keyword**	**Sum**	**Keyword**	**Sum**
Event (K)	*Keluhan*	10000	*Obat*	10000	*Setelah minum obat*	65	*Obat tidur*	1167	*Obat batuk pilek*	38
Drug(O)	*Amlodipin*	3	*Kapsul*	2900	*Puyer*	251	*Serbuk*	3500	*CTM*	470
Effect(E)	*Muntah*	10000	*Ngantuk*	10000	*Muntah*	10000	*Gatal*	8904	*Pusing*	10000
Food (M)	*Sprite*	797	*Semangka*	4200	*Buah*	10000	*Jus*	5381	*Es krim*	6300
	TOTAL	**20800**	**TOTAL**	**27100**	**TOTAL**	**20316**	**TOTAL**	**18952**	**TOTAL**	**16808**

Table 6. Filtering Process Result

Case no	Event	Drug	Effect	Food	Total
1	*Keluhan*	*Amlodipin*	*Muntah*	*Sprite*	
Sum	0	0	2	3	**5**
2	*Obat*	*Kapsul*	*Ngantuk*	*Semangka*	
Sum	111	651	67	23	**852**
3	*Setelah Minum obat*	*Puyer*	*Muntah*	*Buah*	
Sum	1	1	23	5	**30**
4	*Obat tidur*	*Serbuk*	*Gatal*	*Jus*	
Sum	5	117	293	5	**420**
5	*Obat batuk pilek*	*CTM*	*Pusing*	*Es krim*	
Sum	3	1	7	5	**16**
				TOTAL	**1.323**

IV. CONCLUSION

This study has explored data collection based on specific keywords for drug and food interactions. The collection media used is the Twitter API. There are two scenarios tested in this study. The first test scenario is data collection based on a combination of two or more keywords. The second test is data collection using a single keyword search mechanism and filtering process. The results showed that testing with the best keyword search mechanism is a single keyword and the process of filtering the amount of data collected as much as 1.332 data. This amount is higher than the data collection mechanism using a combination of two or more keywords that only collected as many as 438 data.

ACKNOWLEDGMENT

The research leading to these results has received funding from the Ministry of Research Technology and Higher Education Indonesia.

REFERENCES

[1] F. Rahmawati, N. Hidayati, W. Rochmah, and S. A. S. Sulaiman, "Potentiality of drug-drug interactions in hospitalized geriatric patients in a private hospital, Yogyakarta, Indonesia," Asian J. Pharm. Clin. Res., vol. 3, no. 3, pp. 191–194, 2010.

[2] C. Paris, H. Christensen, and P. Batterham, "Exploring emotions in social media," no. 2013, 2015.

[3] A. Sriteja, "Controversy Detection Using Reactions on Social Media," 2017.

[4] H. Khotimah, "Pemodelan hybrid tourism recommendation menggunakan hidden markov model dan text mining berbasis data sosial media husnul khotimah," 2014.

[5] K. Chen, T. Chen, G. Zheng, O. Jin, E. Yao, and Y. Yu, "Collaborative personalized tweet recommendation," Proc. 35th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. - SIGIR '12, no. 800, p. 661, 2012.

[6] "1 Second - Internet Live Stats," Internet Live Stats. [Online]. Available: http://www.internetlivestats.com/one-second/. [Accessed: 24-Apr-2018].

[7] S. Lomborg and A. Bechmann, "Using APIs for Data Collection on Social Media," Inf. Soc., vol. 30, no. 4, pp. 256–265, 2014.

[8] Twitter, "Standard search - Twitter Developers," Twitter Developers, 2018. [Online]. Available: https://developer.twitter.com/en/docs/tweets/search/overview/standard. [Accessed: 25-Apr-2018].

[9] Stieglitz, Stefan; Dang-Xuan, Linh; 2013; "Social media and political communication: a social media analytics framework, Social Network Analysis and Mining", Volume 3 Issue 4, 1277-1291

[10] Ivana Grujic; Sanja Bogdanovic-Dinic; Leonid Stoimenov; 2014; "Collecting and Analyzing Data from E-Government Facebook Pages; ICT Innovations 2014 Web; 86-96

[11] Liang, Tingting;Chen, Liang;Ying, Haochao;Zheng, Zibin; Wu, Jian; 2016; Crowdsourcing based API Search via Leveraging Twitter Lists Information; Proceedings - 15th IEEE International Conference on Data Mining Workshop, ICDMW 2015, 1540-1547

[12] Emilio Ferrara; Pasquale De-Meob; Giacomo Fiumara; Robert Baumgartner;2014;Web data extraction, applications and techniques: A survey; Knowledge-Based Systems Volume 70; 301-323

[13] Chang, Chun Hao;Saravia, Elvis;Chen, Yi Shin; 2016;Subconscious Crowdsourcing: A feasible data collection mechanism for mental disorder detection on social media;Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016; 374-379;

[14] Gentry, Jeff, 2016, "R Based Twitter Client", CRAN Project[Online]. August 29, 2016. Available: https://cran.r-project.org/web/packages/twitteR/twitteR.pdf;; accessed : 25-Apr-2018

[15] Twitter, "Standard API Rate limits per window", https://developer.twitter.com/en/docs/basics/rate-limits, [Aceessed: 28 April 2018]

Ultranus: a Novel Indonesian Cultural Game using Artificial Intelligence

Adelia Rokhmawati, Gradiyanto Radityo Kusumo, Irawan Dwi Wahyoho
Department of Electrical Engineering
Universitas Negeri Malang
Malang, Indonesia
Email : adelia.rokhmawati.1505336@students.um.ac.id,
gradiyanto.radityo.1405334@students.um.ac.id,
irawan.dwi.ft@um.ac.id,

Rofiana Irawati
Department of Art & Design
Universitas Negeri Malang
Malang, Indonesia
Email : rofiana.irawati.1502536@students.um.ac.id

***Abstract*— The development of smartphone technology so rapidly as well as the development of games on smartphones. The interest of the game users on the smartphone depends on the criteria and features of the game that is the category and difficulty level in the game. Interest and level of children's games are different from adult users. This research develops a snake ladder game called Ultranus that is for children's games and children with special needs. This game uses snakes and ladders as a storyboard and rule. Storyboard on Indonesian culture. Determination of categories and levels in this game using Artificial Intelligence. Level determination using forward chaining algorithm while category determination using backward chaining algorithm. The results of the implementation and testing of this game obtained level of accuracy for the level of 80% and the category of 90%.**

Keywords—Game, Artificial Intelligence, Forward Chaining, Backward Chaining.

I. INTRODUCTION

The development of information technology is currently very rapid, especially for smartphones and computers. The development of smartphone technology in Indonesia is increasing rapidly from year to year, which is 17-20% increase every year, this proves that technology is very influential on Indonesian society, especially children [1-2]. Technological developments also affect children's activities. It indicated by the increasing number of a phenomenon of children who hold a smartphone every day. This phenomenon reinforced with children who prefer a smartphone than playing with peers. For that required education facilities through a smartphone in the form of games and entertainment according to their age level [4-5].

Since 2013 in Indonesia, 72 percent of children under the age of 8 have used gadgets such as smartphones, tablets, and iPods, where the majority of 2-year-olds prefer to use tablets or smartphones every day. Compared to 2011 the figure is still at 38 percent [6]. As shown in Figure 1 where the survey results in 2015 show 55 respondents smartphone users aged 9-13 years with various brands of smartphones used by elementary school students grade 4, 5, up to 6 and indirectly also indicate that the respondent has been bought smartphone by their parents, this survey was conducted by a teacher who was in one of the primary schools with the category of parents of the middle-class economy [7].

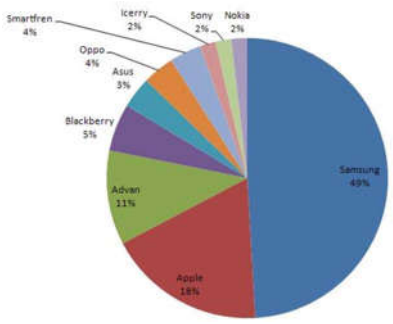


Fig. 1. Smartphone User Data

One of the content favored by kids smartphone users is games. The game often accused of having a negative influence on children. Games have positive functions and benefits for children, among them, children know computer technology, lessons to follow directions and rules, problem-solving and logic exercises, train motor nerves and spatial skills, build parental communication while playing together, and provide entertainment [7]. However, it is still rare to find educative games for children, especially those that fit the category and level of difficulty. Games that are supportive of learning in schools, especially education culture Indonesia is still challenging to find.

Snake ladder is one of the traditional games in Indonesia. The ladder snake is made up of boards and stairs that are matching by 2 or more people. The ladder snake game is a square-shaped board game that divided into small squares, and on each box, there are two possibilities: a ladder or a snake that will connect with the other boxes. This game made in 1870. There is no standard board game on the landing snakes; everyone can make his or her boards with different boxes, snakes, and ladders [8].

This research develops smartphone application based on Expert System. This application is an audio-video educational game that aims to improve children's understanding of the culture of the Indonesian archipelago. In this game has the selection of levels and categories using Expert System so that

users feel comfortable playing the game by the level of user ability that is children and children special needs.

II. ARTIFICIAL INTELLIGENCE

A. *Selection level and category*

The game designed in this study is a leveling game, where the power of the game lies at different difficulty levels at each level. This game runs android platform on a smartphone. In one game this game can be played by the user and multiuser.

In this Ultranus Game utilize Artificial Intelligence (AI). AI used is FC (Forward Chaining) method and BC (Backward Chaining) method. Mainly the forward chaining method is used to determine whether the user can continue the game to the next level or not. The forward chaining method is an algorithm that focuses on a data-driven approach. In this approach, the tracking starts from the input information and then tries to illustrate the conclusions. In this case, it is called the inference multiplication chain that connects the problem with its solution. The forward chaining method, ie the chain is searched or traced from the problem to get the solution [9]

According to Meenakshi Batra and Vijay Batra in his journal in March 2006 stated that the concept of forwarding Chaining is useful enough to train mentally disabled children [10]. In this ultranus game using backward chaining method to determine the category of this game. The ultranus game category divided into four categories for children with classification, i.e., 1-3 years, 4-7 years, 8-10 years and special needs.

B. *Forward chaining*

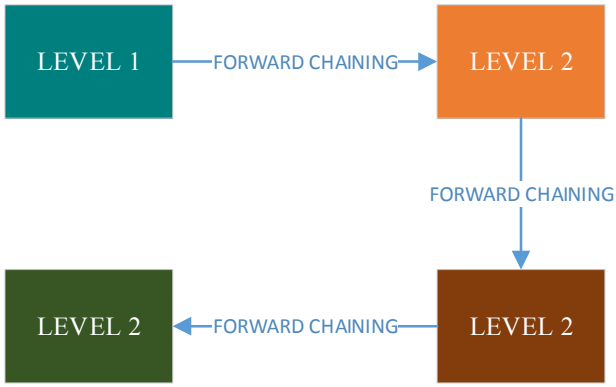


Fig. 2. Rule of the level game

The Forward Chaining method is a chain that follows from the problem of getting a solution. Another way of describing forward chaining is to refer to facts to conclude. [11]. The forward chaining method also defined as a data-driven approach. In this tracking, approach begins with information input, and then try to illustrate conclusion. So this method is also often called "Data-driven"[11].

In carrying forward chaining methods, there are several steps: first, incorporating a known set of facts into working memory; second, deriving new facts according to the rules by a place with known facts. In the forward chaining method, this process continues until it reaches the goal or no more extended rules that match the known facts. So the path taken for the Steps that create the rule based on forwarding chaining is [11]:

1. The problem. In this process is to determine the problem domain and knowledge acquisition.
2. Input Data. The next process is to enter the initial data to start the inference.
3. Data Control Structure. In a complex game or application model requires an extra premise to help control.

The use of variables in this game as a method of forwarding chaining method for the development of this game is as described in Table 1. The result variables used as described in Table 2 and the level variables used as described in Table 3. The process of determining the level using forward chaining as described in Figure 2.

TABLE I. VARIABLE RULE.

No.	Rule Code	Condition	Information
1.	R01	[0,1,2,3]	Bonus
2	RXX	....	.....

TABLE II. VARIABLE RESULTS

No.	Rule Code	Review
1	H01	Congratulations, Gold, Coin, Bonus
2.	HXX	....

TABLE III. VARIABLE LEVEL

No.	Rule Code	Information
1	L01	Next Level
2	LXX	....

C. *Backward chaining*

Backward Chaining is reasoning that starts from the hypothesis and then traces back the facts that support the hypothesis. Backward chaining also described as the reasoning form from goal to subgoal with the understanding of reaching a goal means to meet the subgoal therefore backward chaining reasoning is also called top-down reasoning [12]. The inference chain for backward chaining can see in Figure 3, where the reasoning process starts from H1 or H2 or H3, suppose we start from H1 it will look for facts related to H1 to prove that hypothesis H1 is correct. The value of fact A is unknown then H1 cannot be verified, then the inference process switches to H2. Moreover, so on until obtained evidence/evidence that proves the truth of a hypothesis[12].

In designing this game for the required category of facts as shown in table 4 which contains about 60 questions randomly assigned as much as 5 for each category with the number of categories as five shown in table 5. After determining the facts and reasoning process then make a decision tree like that shown in Figure 3.

TABLE IV. QUESTIONS ABOUT CATEGORIES

No	Question	A	B	C	D	E
1	Is Mr. Jokowi the President now?	B	B	B	B	B
2	Next Question	...	...	...	...	...

TABLE V. CATEGORY

No	Category	Information
1	A	Age 3-6
2	B	Age 7-9
3	C	Age 10-12
4	D	Children Special Needs Age 3-8 years
5	E	Children Special Needs aged 9-12 years

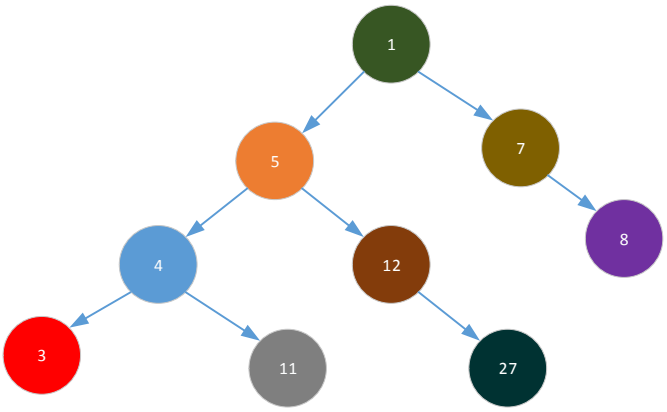


Fig. 3. Tree of decision

III. DESIGN SYSTEM

In this section is divided into two namely the method used and the user interface in this game. The method used in designing and implementing this game by the algorithm method that will use is forward chaining and backward chaining. For the user interface, the selection of user interfaces based on data collection from media experts, model experts, and material experts for children's games. The user interface option also based on a child with special needs.

A. Method

This study used the experimental method [13] as shown in Figure 4. This study begins with the study of interest literature and user model of the game in android and then collecting data from snake ladder game. After the two steps above is done algorithm design to perform user classification in the form of category and level of difficulty game in the form of level. The user category uses backward chaining algorithms and level determination using forward chaining algorithms. The next step is to design the game as a whole and then check the accuracy for categories and levels. If the accuracy of the game is more than 70%, then the game is uploaded play store.

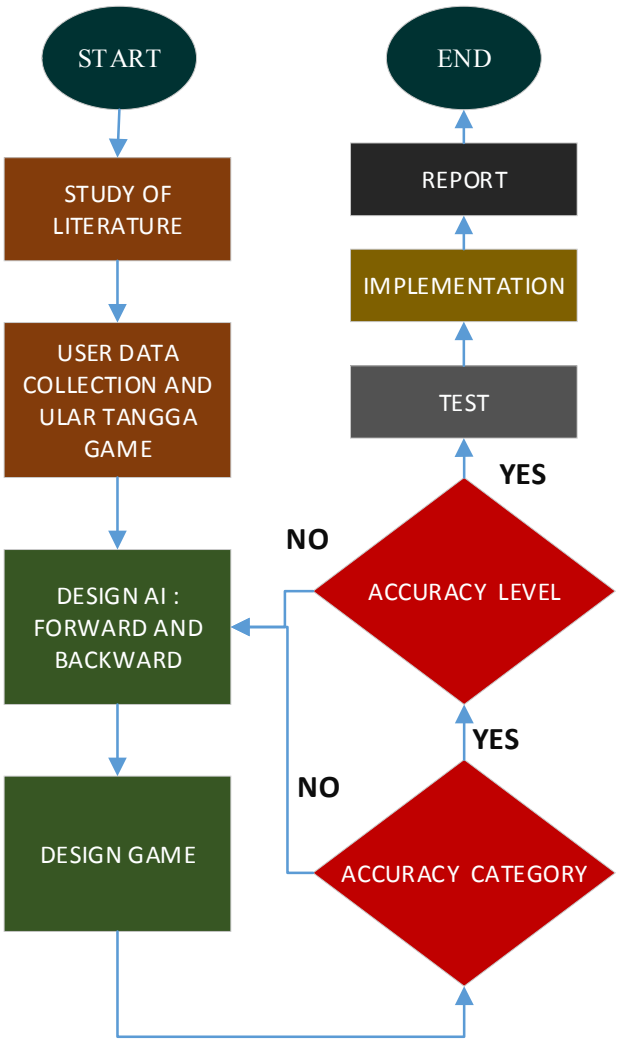


Fig. 4. Method

B. User Interface

The Ultranus Game view is divided into several views as shown in Figure 5, namely:

1. The initial game view as shown in Figure 5 (a) where the user interface shows the theme of this game.



Fig. 5. Interface of the ultranus game (a) The initial game, (b) Game start choice, (c) Multiplayer choice, (d) Carakter choice, (e) Start Game, (f) Cultural Knowledge, (g) End Game and (h) Choice for Play again

2. Initial game view by giving the option to start the game or exit as shown in Figure 5.b
3. Display options, there are two options namely single or more than one as shown in Figure 5.c.
4. Display the player character selection to used as shown in Figure 5.d
5. The initial view starts the game. The selection is a single user where the user fights his smartphone as shown in Figure 5.e.
6. Display where each in-game box has its feature with the Indonesian culture storyboard as shown in Figure 5.f.
7. Display Finish the game, where the game has been completed and won by either party as shown in Figure 5.g
8. End Display game. In this view, the user chooses to continue the game or finish the game as shown in Figure 5.h.

IV. RESULT

Testing does testing [14] in this game at the level and category.

A. Level

Testing [9] at the level based on the AI method used forward chaining. Test data done by looking at the level of user success gained when users download the level and register their account online on Tuesday, April 17, 2018, 21.30.40. Based on the database of registered users, time created and download of the next level scheme as shown in Table 6.

TABLE VI. DATABASE ACCOUNT

No	User/ Alias	Registrasi	Created	Download
1	Usr0 001	Tuesday, 17 April 2018, 21.30.41	Wednesday, 11 April 2018, 08.45.34	Wednesday, 11 April 2018, 08.45.34
2	Usr0 002	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41
3	Usr0 003	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41
4	Usr0 004	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41
5	Usr0 005	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41
6	Usr0 006	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41	No Data
7	Usr0 007	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41
8	Usr0 008	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41
9	Usr0 009	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41	No Data
10	Usr0 010	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41	Tuesday, 17 April 2018, 21.30.41

B. Category

Testing [9] on the category based on AI method used is backward chaining. Test data is obtained from user feedback when the game finishes and the level of user satisfaction. At the time accessed Tuesday, April 17, 2018, 21.30.41. The category conformity parameter is supported by downloading the next level as shown in Table 7.

TABLE VII. DATABASE ACCOUNT CATEGORY

No	User/ Alias	Accessed	Category	Download
1	Usr00 01	Tuesday, 17 April 2018, 21.30.41	suitable	Wednesday, 11 April 2018, 08.45.34
2	Usr00 02	Tuesday, 17 April 2018, 21.30.41	suitable	Tuesday, 17 April 2018, 21.30.41
3	Usr00 03	Tuesday, 17 April 2018, 21.30.41	suitable	Tuesday, 17 April 2018, 21.30.41
4	Usr00 04	Tuesday, 17 April 2018, 21.30.41	suitable	Tuesday, 17 April 2018, 21.30.41
5	Usr00 05	Tuesday, 17 April 2018, 21.30.41	Sesuai	Tuesday, 17 April 2018, 21.30.41
6	Usr00 06	Tuesday, 17 April 2018, 21.30.41	No	No Data
7	Usr00 07	Tuesday, 17 April 2018, 21.30.41	Sesuai	Tuesday, 17 April 2018, 21.30.41
8	Usr00 08	Tuesday, 17 April 2018, 21.30.41	suitable	Tuesday, 17 April 2018, 21.30.41
9	Usr00 09	Tuesday, 17 April 2018, 21.30.41	suitable	No Data
10	Usr00 10	Tuesday, 17 April 2018, 21.30.41	suitable	Tuesday, 17 April 2018, 21.30.41

V. DISCUSSION

Discussion of test results from this game app [9] based on level and category. The results of this test to see the level of accuracy of game level determination based on forwarding chaining algorithm and game category determination based on backward chaining algorithm.

A. Level

The discussion for this level of accuracy I based on ten accounts accessing this app and downloading this app according to table 5 is 80% where two accounts do not continue downloading the next level scheme. Causes of a failed user where the user is not continuing this game, or the level of difficulty of this level does not match the user.

B. Category

The discussion for the accuracy of this category based on the ten accounts that access this app and provide feedback as shown in table 6 is 90% where there is only one account that provides feedback does not match the category of users and does not keep downloading the next level. However, there is one account that provides feedback by the category but does not continue downloading. One user failed to provide feedback due to where the user did not continue this game, or the level of difficulty of this level did not match the user.

VI. CONCLUSION

This research develops ultranus game that is game of snake ladder on the android system in a smartphone with the determination of level and user category using AI (Artificial Intelligence). The Ultranus game category divided into four categories for children with classification, i.e. age 1-3 years, age 4-7 years, age 8-10 years and special needs. Game category determination using backward chaining algorithm. In this game, a game level determination is level 1-5 using forward chaining algorithm. The results of testing in this game obtained that the accuracy of determining the level of 80% and determination category by 90%.

ACKNOWLEDGMENT

This research supported by PKM 2018 fund from the Directorate General of Learning and Student Affairs at Ministry of Research, Technology and Higher Education of the Republic of Indonesia.

REFERENCES

[1] Samaha, M., & Hawi, N. S. (2016). Relationships among smartphone addiction, stress, academic performance, and satisfaction with life. *Computers in Human Behavior*, *57*, 321-325.

[2] Chiu, S. I., Lee, J. Z., & Huang, D. H. (2004). Video game addiction in children and teenagers in Taiwan. *CyberPsychology & behavior*, *7*(5), 571-581.

[3] Kwon, M., Kim, D. J., Cho, H., & Yang, S. (2013). The smartphone addiction scale: development and validation of a short version for adolescents. *PloS one*, *8*(12), e83558.

[4] Lee, S. J., Kim, B., Choi, T. K., Lee, S. H., & Yook, K. H. (2014). Associations between smartphone addiction proneness and psychopathology. *Korean Journal of Biological Psychiatry*, *21*(4), 161-167.

[5] Mok, J. Y., Choi, S. W., Kim, D. J., Choi, J. S., Lee, J., Ahn, H., ... & Song, W. Y. (2014). Latent class analysis on internet and smartphone addiction in college students. *Neuropsychiatric disease and treatment*, *10*, 817.

[6] Oliver, E. (2010, June). The challenges in large-scale smartphone user studies. In *Proceedings of the 2nd ACM International Workshop on Hot Topics in Planet-scale Measurement* (p. 5). ACM..

[7] Chen, C., Zhang, K. Z., Zhao, S. J., Lee, M. K., & Cong, T. (2016, January). The Impact of Mere Exposure Effect on Smartphone Addiction. In *System Sciences (HICSS), 2016 49th Hawaii International Conference on* (pp. 1507-1514). IEEE.

[8] Sheth, J. R. (2014, July). Snake and ladder based algorithm for steganographic application of specific streamline bits on prime gap method. In *Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on* (pp. 12-16). IEEE.

[9] Durkin, J., & Durkin, J. (1998). *Expert systems: design and development*. Prentice Hall PTR.

[10] Batra, M., & Batra, V. (2006). Comparison between forward chaining and backward chaining techniques in children with mental retardation. *The Indian Journal of Occupational Therapy*, *37*(3), 57-63.

[11] Slocum, S. K., & Tiger, J. H. (2011). An assessment of the efficiency of and child preference for forward and backward chaining. *Journal of applied behavior analysis*, *44*(4), 793-805.

[12] Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited.

[13] Basili, V. R. (1993). The experimental paradigm in software engineering. In Experimental Software Engineering Issues: Critical Assessment and Future Directions (pp. 1-12). Springer, Berlin, Heidelberg.

[14] Stuart-Fox, D. (2006). Testing game theory models: fighting ability and decision rules in chameleon contests. *Proceedings of the Royal Society of London B: Biological Sciences*, *273*(1593), 1555-1561.

[15] Suwasono, S., Prihanto, D., Wahyono, I. D., & Nafalski, A. (2017). Virtual Laboratory for Line Follower Robot Competition. *International Journal of Electrical and Computer Engineering (IJECE)*, *7*(4), 2253-2260.

[16] Wirawan, I. M., Taufani, A. R., Wahyono, I. D., & Fadlika, I. (2017, October). Online judging system for programming contest using UM framework. In *Information Technology, Computer, and Electrical Engineering (ICITACEE), 2017 4th International Conference on* (pp. 230-234). IEEE.

[17] Wirawan, I. M., Herwanto, H. W., & Wahyono, I. D. (2017, October). Virtual programming laboratory UM (VLPUM) for programming competition. In *Electrical, Electronics and Information Engineering (ICEEIE), 2017 5th International Conference on* (pp. 158-162). IEEE.

Outpatient Registration System Based On Android At BBKPM Surakarta

Fahrindra Yusuf Arfannedy
Medical records and health information Department,
Medical records and health information Academy
APIKES Citra Medika
Surakarta, Indonesia
fahrindra.yusuf19@gmail.com

Tominanto
Medical records and health information Department,
Medical records and health information Academy
APIKES Citra Medika
Surakarta, Indonesia
tommynanto@gmail.com

Andi Yulianto
Medical records and health information Department,
Medical records and health information Academy
APIKES Citra Medika
Surakarta, Indonesia
Yulianto_andi@rocketmail.com

***Abstract*- The development of information and communication technology increases rapidly, including the development of mobile phone technology that brings smartphone with various functions, one of them is as a source of information. It can encourage the public to access information and interact quickly. Along with these developments, BBKPM Surakarta has implemented a Short Message Service based registration system, but the system still has problems, such as Short Message Service replies received by patients is not appropriate. The strategy undertaken in accelerating the service is the outpatient registration system based on the android at BBKPM Surakarta. The purpose of this research is to make an outpatient registration system based on android at BBKM Surakarta. Type of the research uses descriptive, identification of variables used are patient data, physician data, polyclinic data, and payment data. The subject in this research is the registration officer, and the object in this research is part of registration and IT. Instruments used in this study are observation and interview, as well as used System Development Life Cycle (SDLC) system development method. The outpatient registration system based on Android at BBKPM Surakarta, was built to accelerate the process of service, especially in the registration section, with this application is expected that later patients can register anytime and anywhere, and can see the doctor's schedule before applying through an existing application on Android. Hospital information system at BBKPM Surakarta should be developed better, so that health service given to society will be more efficient to process of health service at the hospital.**

Keywords- Registration System; Outpatient; Android

I. INTRODUCTION

The world of information technology that is growing has brought a rapid influence in health field, one of them is internet. The Internet is a computer network that can provide information more quickly and efficiently to the community. Internet technology embedded in various devices such as mobile phones. In this era, technological development of mobile phones has not only a communication function but also access the internet site. This is in use by some mobile manufacturers in creating a smart phone that is more known as 'Smartphone'. Many mobile phone manufacturers create smartphones that have sophisticated features that are needed by community communication. One of the Operating System (OS) on smartphones that are growing rapidly is Android. Many people who use android, ranging from small children to adults. This is due to the increasing number of applications and games that are circulating on the android.

Android users start to enter the health world as well, the world is hospital. With the information technology development, it can support the hospital service faster, more accurate and more efficient so that it can improve the hospital's quality.

The health service institution for communities with their own characteristics are influenced by the development of health sciences, technological advancement, and socio-economic life. The health service institutions must be able to improve the quality of services and able to be affordable by the community to realize the highest level of health.[1]

use of a computerized system, is expected to improvequality of administration, [2] a computerized system is very necessary, because the services provided at the Puskesmas must also be fast. For example, addressing the management information system for the registration of patients who have been used. [3] BBKPM Surakarta is one of the pulmonary health service facilities located in Surakarta. BBKPM Surakarta was established in 1957 under the name of Balai Pemberantas Penyakit Paru-Paru (BP4) Surakarta with service at that time addressed to patient of Pulmonary TB only. Because of the issue of Permenkes 2354 / MENKES / PER / XI / 2011, In 2011 it renamed into the Community of Lung Health Center (BBKPM) Surakarta. The existence of BBKPM Surakarta is under the Directorate General of Health and technically functioned by the Directorate of Health Care Referrals.

BBKPM Surakarta has 7 polyclinics with an average visit of 200 patients per day coming from various regions in around Solo, so the registration takes long queue. Out patient registration system at BBKPM Surakarta has been done computerized. New patient registration is done directly at the hospital, and registration of old patient can be done through SMS (Short Message Service). Registration via SMS is done by patients with the format in accordance with the predetermined and sent to the number that has been determined, then the data is entered into the server and the patient gets a message. With the registration via SMS has many advantages, but in addition it also has some disadvantages, such as if the patient mistyped the SMS format then the data cannot enter the server, then the officer must check the server pulse on a regular basis, because when the SMS pulses on the server run out then the patient will not get SMS replies.

II. THEORETICAL FOUNDATION

A. Medical Record

Medical records is a file containing records and documents on the patient's identity, examination, treatment,

actions and other services that have been provided to the patients.[4]

The purpose of medical records is to support the achievement of orderly administration in order to improve health services at hospitals. Without a proper medical records management system and properly administered hospital administration will not work as expected. While the administrative order is one of the factors that determine the efforts of health services in hospitals [5]

B. Health Information System

Health information systems are a support for implementation good health management in support of improved performance health services. to realize this, health information must be used as a material for consideration for decision makers at each level of management health. [6]

C. Outpatient Acceptance Site (TPPRJ)

Outpatient Acceptance Places or Outpatient Registration Sites (TPPRJ) are also called Outpatient Registration Counters. [7]

D. System Design

System design is a complementary problem-solving technique (with system analysis) that reassembles component parts into a complete, improved system of expectations. This involves the addition, deletion, and relative parts changes in the original system (original). In designing process models and data models to realize the defined system requirements and the framework for coding is also specified. Documentation of this design stage is very useful for future system development, if there is a change of user needs. [8]

E. System Development

System development means setting up a new system to replace the old system as a whole or improve the existing system.[9]

III. MEDHODS

A. Types Of Research

The research method used is descriptive with survey, the method uses to assess the condition and the implementation of a program in the present, then the result is used to plan the improvement program. The approach used is cross sectional that is a reset to study the dynamics of correlation between risk factors with effects, by approach, observation or data collection at one time. [10]

B. Identification of variables

Identification of variables used by the author in the preparation of scientific papers are: (1) outpatient procedures and grooves (2) Patient's data (3) Doctor's data (4) Payment Data (5) Polyclinics data

C. Research Instrument

Guidelines for interviews with the submission of questions are either orally or in writing to obtain information. The guidelines in the study are used to obtain information from TPPRJ officers in BBKPM Surakarta.

D. Data Processing Research

Data collection is data collection from hospital medical records unit consisting of BBKPM Surakarta profile data, outpatient form, BBKPM Surakarta protocol, patient's data, polyclinics information data.

Editing is correcting the data that has been collected before being processed because the possible data still duplicate data and errors.

Classification is the classification or differentiation of data that has been obtained in accordance the classification into table form such as patient's table, polyclinics table, and registration table. Presentation of data is data that has been collected and they will be processed with PHP programming language and MySQL database which will generate output in the form of report of patient visit, Report on Patient Visits Per Payment Method , report per period, and report of information of polyclinics, and proof of online registration**.**

E. System Development Methodology

- Project Identification and Selection, At this stage, several things to be done are to identify potential projects, classify and rank projects, and select projects to develop. Researchers conducted an initial survey to determine the problems that exist in BBKPM Surakarta.
- Initiation And Project Design, At this stage a potential registration system project is described and arguments for continuing the project are put forward. Researchers plan to build an outpatient's registration system based on android in bbkpm Surakarta. At this stage the researchers also conducted literature studies and research studies. The literature study aims to understand concepts and materials related to outpatient information systems to support the completeness of the theory and understanding of the concept of the cases at hand. Previous research studies on outpatient information systems to improve understanding and comparison of authenticity of research.
- The analytical stage is the stage at which the system is currently being studied and a replacement system is proposed. This stage researchers study the old system in a way, there are: Gathering information and Defines system requirements.
- The design stage is the phase change needs that still form the concept into a real system specification. Stages of system design can be divided into 2 stages, namely logical design (logical design) and physical design stages (physical design). At this stage the researchers design the system specifications to be built include: (1) Designing the flow of data with flowchart (2) Designing Entity Relationship Diagram (ERD) and Data Flow Diagram (DFD) (3) Designing input, this design consists of several input data, among others: (a) Patients data (b) Polyclinics data (c) Check date (d) Payment Data (4) Designing process that will be used to design outpatient information system using PHP programming language with MySQL database and Dreamweaver software. (5) Designing Output, this design consists of several outputs, among others: (a) Online Registration Proof (b) Patient Visa Report (c) Report on Patient Visits Per Payment Method (d) Report on

Patient Visits Per Period (e) Report of Patient Visits Per Doctor (f) Online Patient Visa Report

- Impementation stage the researchers test the resulting program code and can be applied to the project to be developed, after testing the program then the program will be installed at BBKPM Surakarta to replace the old system.
- The final step of the SDLC where at this stage the system is systematically improved and improved. The result of this stage is a new version of software that has been created. Improvements made at the level can be very varied, ranging from repairing programs that crash to work again until the addition of new program modules in response to changing needs.

IV. RESULT

Place of Outpatient Registration (TPPRJ) BBKPM Surakarta, there are 2 types of registration that are direct registration and registration via SMS **(Short Message Service)**. Direct registration is the patient came to BBKPM Surakarta and directly register at the registration to be entered by the registration officer. Registration at BBKPM already using computerization. Registration via SMS is done with the patient register via SMS in accordance with predefined format. Patients who register via SMS, maximum 1 day before checking and are at work hours at 08.00-16.00 (the officer must check the credit, sometimes inappropriate sms replies, queue). Here is the following system at BBKPM Surakarta:

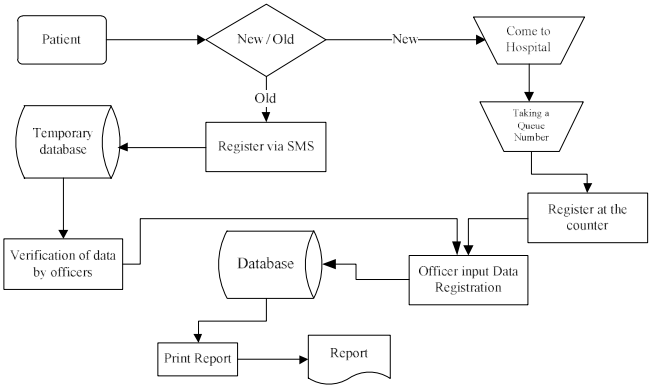


Fig. 1. Current System at BBKPM Surakarta

The system developed at BBKPM Surakarta is outpatient registration based on android. The following chart system of outpatient registration is based on android to be developed at BBKPM Surakarta.

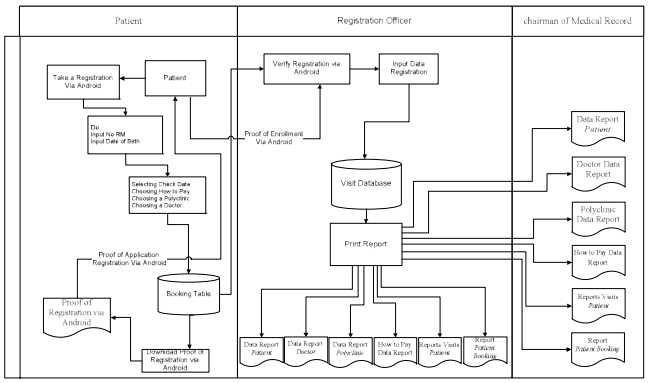


Fig. 2. The system built at BBKPM Surakarta

A. System Planing

1) Data Flow diagrams (DFD) consists of context diagrams, tiered diagrams, DFD level 0 and DFD level 1. (1) Diagram korteks :

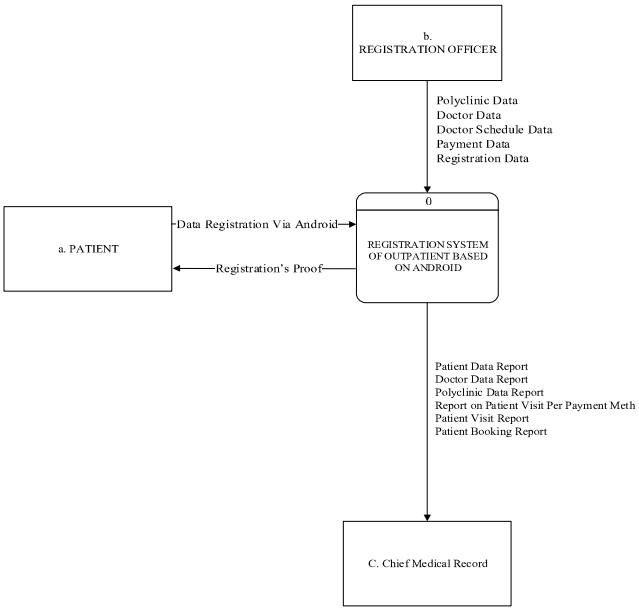


Fig. 3. Context Diagram

2) *DFD level 0:*

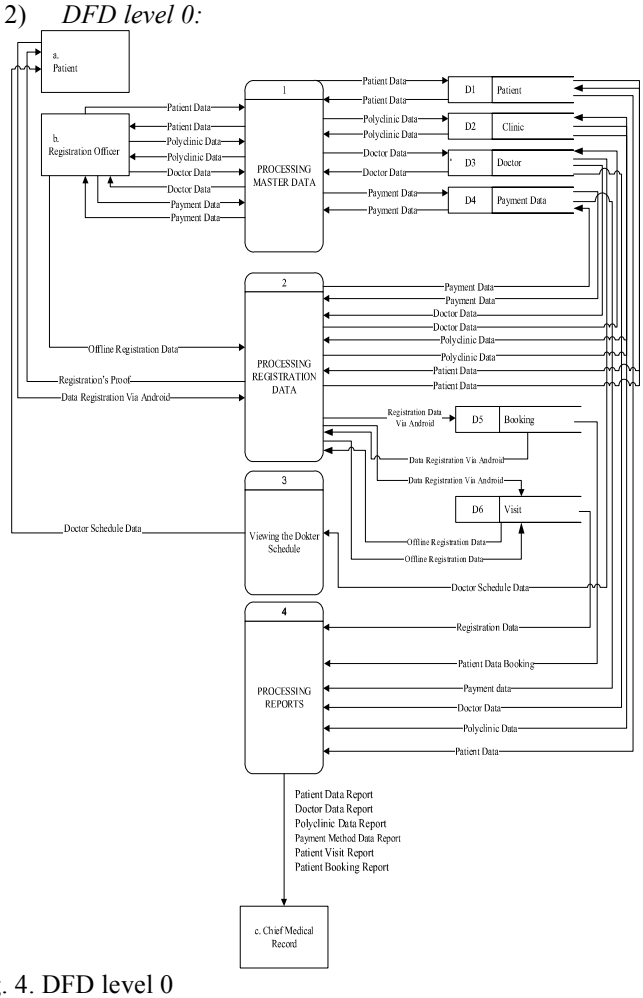


Fig. 4. DFD level 0

B. Table Relation

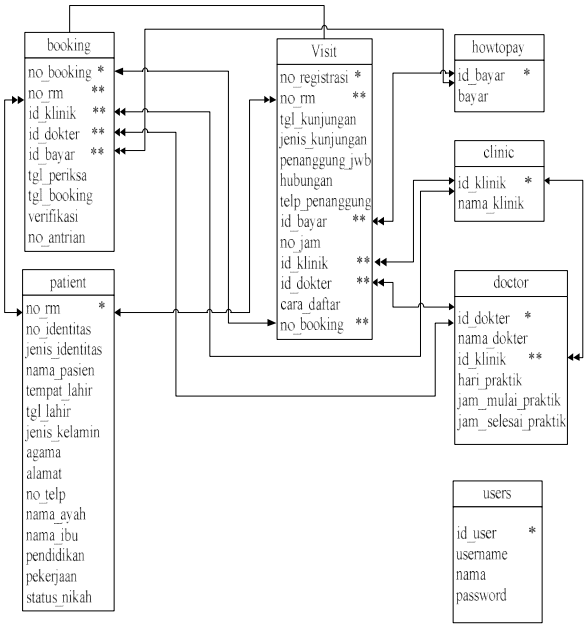


Fig. 5. DFD Relation table

C. How to Operate Program

a) Android users

(1) The activation of the program starts with running the app on Android,Once the application is successfully executed, it will show the main page. The main page of this application has the name of the hospital, and in the top left corner there is a menu.



Fig. 6. Program Activation On Android Users

(2) The options menu is in the top left corner, when clicked it will appear several options, among others: Home, Doctor Schedule, Outpatient Registration, and Location.



Fig. 7. Options Menu

(3) The doctor's schedule serves to see the doctor's schedule including doctor's name, practice day, start and end time of practice.

(4) The outpatient registration menu serves to perform an outpatient registration that can be performed by an elderly patient or patient who has previously been treated at BBKPM Surakarta.



Fig. 8. Patient Data Validation

(5) Validate patient data using medical record number and patient's birth date. If the patient is wrong in the input, medical record number and date of birth it will appear notification that patient data is not found, and if true it will show patient data. In this stage the phone number can be changed when there is a change in patient data. select next at the bottom so that the display appears.

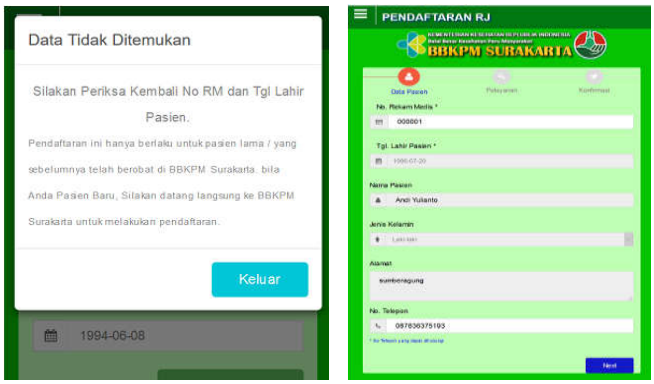


Fig. 9. Patient Data Notification Not Found and Data Patient

(6) At this stage the patient can choose the date of check, how to pay, clinic purpose, and the name of the doctor. Selected check dates are three days after registration is made and a maximum of one day before check. After all the complete data, then press the verification button, if the patient had previously done registration, but not yet come to BBKPM Surakarta then do re-registration, it will show the notification.



Fig. 10. Input Data Registration

(7) But if the patient successfully registered then will show the registration notification successfully completed with patient registration data, then the patient can download the proof of registration in the form of pdf files that will be brought when coming to BBKPM Surakarta.



Fig. 11. Registration Successful and Proof of Enrollment

(8) The location menu in this app serves to display the map / location BBKPM Surakarta.

b) Admin User

(1) to activate the administration program is done by visiting the server address such as: http://andi-media.info, the it will appear login page then enter the username dan password.
(2) After successful login, it will appear the main page.
(3) The master data menu serves to manage the master data, on this menu when clicked it will display the Data menu Patient, Polyclinics Data, Doctor Data, and Data Cara Pay.
(4) Master patient data serves manage the patient identity data that has been entered previously. Patient data will be displayed in the table, to search the patient, input the patient's name, and then click the search button. to make data changes can be done by selecting the patient on the table, then click edit and then appear edit data dialog. And to do Clear data, select the patient in the table, then click the delete data button. If data printing will be done, click the print data button located on the top right.
(5) Master of polyclinics data serves manage polyclinics data. To add data, click on the new data button, then input data dialog, input the name of the clinic, and press save button, once data successfully entered it will be displayed in table. To make data changes, select the poly data in the table, then click edit data, it will pop up the edit data dialog, then input the data changes, and click save to update. to perform data deletion, click data in table, then click delete data button, and to print polyclinic data, they can press print button on right side.
(6) Doctor data master serves to manage doctor's data, as well as doctor's schedule on Android app. Doctor data master consists of id, doctor's name, specialist, practice day, and hours of practice. To perform the addition of data, click on new data button, then input data dialog, input the name of doctor, specialist, practice day, and practice hours then press save button, after data successfully entered it will be displayed in table. To make data changes, select doctors data in the table, then click edit data, it will pop up the edit data dialog, then input the data changes, and click save to update. to perform deletion of data, click the data in the table, then click the delete data button, and to print the data the doctor can press the print button on the right.
(7) Master data how to pay serves to manage data on how to pay. To add data, click on new data button, then input data dialog, input type of payment method, n and press save button, after data successfully input will be displayed in table. To change data, select the data in the table, then click edit data, it will pop up the edit data dialog, then input the data changes, and click save to update. to perform deletion of data, click the data in the table, then click the delete data button, and to print data how to pay they can press the print button on the right.
(8) The Registration menu is used to perform patient registration transactions, which consists of 2 sub menus, among others, online verification and offline registration.
(9) On this online verification sub menu is used to verify patients who have signed up through android app. Patients who have successfully registered through android application, then the data will be displayed on the booking table to be verified when the patient comes. The medical recorder can print patient data that has been registered via android via the print button, then from the data they can be prepared patient medical records file before the patient comes. To search patient data they can input booking number and name of patient, then click search button, to perform data deletion, select patient on table, then press delete button data, and to do patient verification, select patient at table booking, then press verification button , then the patient verification dialog will appear. Then complete the patient's data, including the name of the person in charge, phone number, how to pay, security card number, the target policlinic, and doctor, then press save button to save the data. The data that has been successfully verified then moved on the table of patient visits. In the table of patient visits, only displays the data of visits on this day.
(10) This sub menu is used to register new and old patients who come directly to BBKPM Surakarta. To register the old patient, click the Offline registration sub-menu, then select the old patient. If the patient is carrying KIB or already known his medical record number, input the medical record number on the available field, if the patient does not bring the KIB or the medical record number is not yet, then search by click the search button to the right of the name field, . input the patient name and address, click select. then the patient data will appear, then complete the registration data, and click save data to save. New patient registration can be done with click of new patient button, then complete patient identity data and visit data. Click save data to save. If the registration is successful, then the patient's identity data will be stored on the pasiaen master, and the visit data will be stored in the visit data.
(11) The report sub menu is used to display report data, consisting of patient data reports, physician data reports, polyclinic data reports, payroll data reports, patient booking reports, and reports of patient visits.
(12) Patient booking report serves to display patient data reports that register through the android application, which will then do the taking of medical record documents from the rack filling before patients come to the hospital. when choosing patient data report order menu then filter data filter will appear based on check date.
(13) The visit report consists of several categories, to view reports of all patient visit data, patient visits per polyclinic, per physician, per way, per period, and per way list, can be

done by clicking the choice of each category then click show, then data will be displayed according to the selected category, to print data with the selected category can be done by clicking the print data button, then the data will be printed.

(14) The access menu consists of user data, which is used to manage user data used for system logins. To add new user data, click on the new data button, then input dialog will show user data, enter the user name and password, then click save to save the data. After the data is successfully saved, then the data will be displayed in the user table. To make data changes, select the user in the table then click the edit button, then it will display the edit data dialog, change data and click save to save. When will do delete data, select data in table then click button delete data.

(15) To exit the app, please click the Logout button on the bottom menu.

V. CONCLUTION

Conclusion of Scientific Writing entitled "System of Registration of Outpatient Patient Based Android at BBKPM Surakarta" that is: (1) Information system of outpatient registration in BBKPM Surakarta is now done with SMS system, but the system still has some weaknesses, such as if patient wrong type sms then data can not enter server, besides officer must check the server pulse periodically, because if the sms pulses on the server run out then the patient will not get sms reply (2) Making information system of outpatient registration based on Android at BBKPM Surakarta, can speed up service process, patient data processing in hospital and more communicative with society about health service in BBKPM Surakarta. The system is designed and applied computerized using the PHP programming language (Hypertext Prepocessor) which consists of patient table, polyclinic table, paying table, physician table, and booking table. The resulting report is patient data report, polyclinic data report, payroll data report, physician data report, patient visit report, per visit doctor report, pay per visit report, visit report per polyclinic, pay per visit report and visit report per period.

REFERENCES

[1]. Undang-Undang No.44 Tahun 2009 Pasal 1

[2]. Akhmad syukron, Noor Hasan, "Perancangan Sistem Informasi Rawat Jalan Berbasis Web Pada Puskesmas Winong" Jurnal Bianglala Informatika Vol 3 No 1 Maret 2015.

[3]. Ana Nur Cahyanti, Bambang Eka Purnama, "Pembangunan Sistem Informasi Manajemen Puskesmas Pakis Baru Nawangan" Journal Speed – Sentra Penelitian Engineering dan Edukasi – Volume 4 No 4 – 2012

[4]. Peraturan Menteri Kesehatan Republik Indonesia Nomor: 269/MENKES/PER/III/2008. Rekam Medis. 12 Maret 2008. Jakarta : Menteri Kesehatan.

[5]. Rustiyanto, Ery. 2011.Sistem Informasi Manajemen Rumah Sakit yang terintegrasi.Yogyakarta: Ardana Media.

[6]. Sudra, Rano Indradi. 2014. *Rekam Medis.* Tangerang Selatan : Universitas Terbuka

[7]. Word Health Organization 2000.

[8]. Sudra, Rano Indradi. 2014. *Rekam Medis.* Tangerang Selatan : Universitas Terbuka.[6]. Fatta, Hanif Al. 2007. *Analisis dan Perancangan Sistem Informasi.* Yogyakarta: Penerbit Andi.

[9]. Jogiyanto (1999) in Yakub. 2012. *Pengantar Sistem Informasi.* Yogyakarta: Graha Ilmu.

[10]. Notoatmodjo, Soekidjo. 2012:35. *Metodologi Penelitian Kesehatan.* Jakarta: Rineka Cipta.

Intelligent System for Agent in Educational Game using Dynamic Gram Similarity

Andhik Ampuh Yunanto, Darlis Herumurti,
Imam Kuswadayan, Siti Rochimah
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
andhikyunanto@gmail.com, darlis@if.its.ac.id,
siti@if.its.ac.id, imam@its.ac.id

***Abstract*—today, almost people like a game. A game can make people have fun when playing it. An exciting game must have many feature and menus especially artificial intelligent (AI) called enemy or agent. Agent in game can be build using any method to make smart and intellect. In educational game, smart agent must can choose and decide a right answer of question. In this research, we make an agent to answer English question using dynamic gram similarity. This method based on n-gram similarity method but it can check neighbor word on sentence. Result of this research indicates dynamic gram similarity has 66% which is more accurate answer than unigram, bigram and trigram similarity. The conclusion of this research is dynamic gram similarity can implement to agent in educational game and can answer grammar English question automatically.**

Keywords— Educational Game, Artificial Intelligent, Text Similarity, Natural Language Processing, Grammar English.

I. INTRODUCTION

Today autonomous agent is used in a lot of domain like commercial and industrial domain [1]. Game is an application in both commercial and industrial domain. Build an agent in game can make student more interested when playing game [2]. Combining education and game can improve educational player skill in a fun way. Because of educational game almost doesn't have agent or intelligent system, a lot of gamer is not playing educational game. So artificial intelligent called agent is important element in game to make more interested for player. Intelligent system for agent in game can build using a lot of method and model such as finite state machine [3], decision tree, rule based system [4], and reinforcement learning [5]. Furthermore, an excitement game must have gamification theorema such as feedback, help, progresive challenge, and reward.

A. Text Simlarity

Similarity is important system for building a heuristic in computational intelligent [6]. According to Han's books, similarity or distance value of neighbor is often used in clustering, prediction, and classification [7]. Nevertheless, other research explain the similarity calculation in text case can using similarity method such as Simple Matching Coefficient (smc) (1), Jaccard Coefficient (jc) (2), and Rao's Coefficient (rc) (3) which is showed in Figure 1. [8].

		X(j)	
X(i)	a	b	
	c	d	

$$S_{smc}\left(x^{(i)}, x^{(j)}\right) = \frac{a+d}{n} \quad (1)$$

$$S_{jc}\left(x^{(i)}, x^{(j)}\right) = \frac{a}{a+b+c} \quad (2)$$

$$S_{rc}\left(x^{(i)}, x^{(j)}\right) = \frac{a}{n} \quad (3)$$

Figure 1. Text Similarity Calculation

B. Text Evaluation

In information retrieval, system performance such as accuracy, complexity, and efficiency is used to evaluation. Furthermore, accuracy is almost evaluated in system intelligent to be known. Figure 2 indicate a formula to calculate accuracy [9]. In comparison of text similarity, n-gram is often used to check similarity between two text or two sentence. N-gram usually used in Markov model to know feature comparison. N-gram method which often used in evaluation is unigram, bigram, and trigram [10]. The differences about three gram method is number text to comparison in each process. Figure 3 is formula of n-gram method.

		Prediction Class	
		Positive	Negative
Actual class	Positive	TP	FN
	Negative	FP	TN

$$acc = \frac{\sum_n \frac{TP+TN}{TP+TN+FP+FN}}{n} \; x\; 100\%$$

Figure 2. Accuracy Formula

$$S = k_1, k_1, \ldots, k_r \quad (5)$$

$$P_n(S) = \prod_{i=1}^{r} P(k_i | k_{i-1}, k_{i-2}, \ldots, k_{i-(n-1)}) \quad (6)$$

Figure 3. n-gram formula

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

C. Our Approach

We assume that intelligent system in game is important to make an interesting educational game. Also it can resolve a boring educational game like question and answer game become more fun. This research focus to build a game education about English education especially for grammar case. Because of previous research concluded that the popular device for educational game in this era is mobile and computer, so we build a game which can be played in computer and mobile device [11]. Furthermore agent was build using heuristic system based on unigram, bigram, and trigram similarity in our previous research [12]. It concluded that agent in game can answer an English question automatically using text similarity. So in this research, we propose a new method for text similarity called dynamic gram. We assume this method can make agent to be better to answer an English question. So it can make game more fun.

II. METHOD

A. Design System Game

Design system in game must have gamification theorem. Based on these theorem, we build a game which has level menus, main menu, battle menu, and result menu. Figure 4 is our game menu.

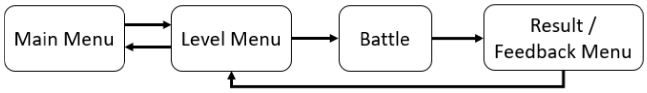


Figure 4. Flow of menus in game

B. Dataset of Question

Dataset for English question in game is a synthetic dataset which is using Longman Toefl Test Books by Philip Deborah. The total data for question is 100 data. This dataset is used to gameplay in game. 100 data of multiple choice is divided in 10 level game.

C. Design Gameplay

Gameplay in this game, player must choose level in level menu. The difficulty level is based on level number. Example, level one until level three is easy level which the English question is same equal with elementary school. Level 4 until level 7 is medium difficulty which is same equal with junior high school. And also the other levels is hard.

After player has already chosen level, battle scene is appearing which like one and one battle. In battle scene, there are one character player called turtle and one enemy called crocodile. Player has to choose the right answer in the middle of question. If player choose the wrong answer, the power attack of player is weak. But if the player choose the right answer, the power is strong and multiplied from normal power. The enemy also must choose right answer. After two both side has already choose the answer, the fastest speed of answer will attack first. After that, the second fastest is attacking the other player. The game will ended if player's health of enemy's health has run out. And then player will go to result or feedback menus. Figure 5 is schema gameplay in battle.

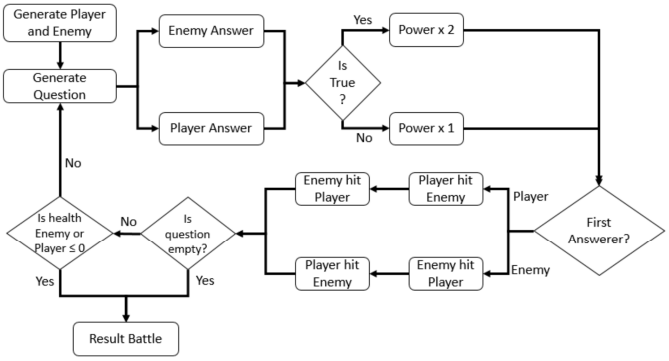


Figure 5. Design gameplay system in battle

D. Design Artificaial Intelligent

Agent in game is using database ground truth for leaning to answer the question. The method is based on similarity which using Simple Matching Coefficient (SMC) and n-gram similarity. The process learning in Agent is just like information retrieval which is showed in Figure 6 and Figure 7.

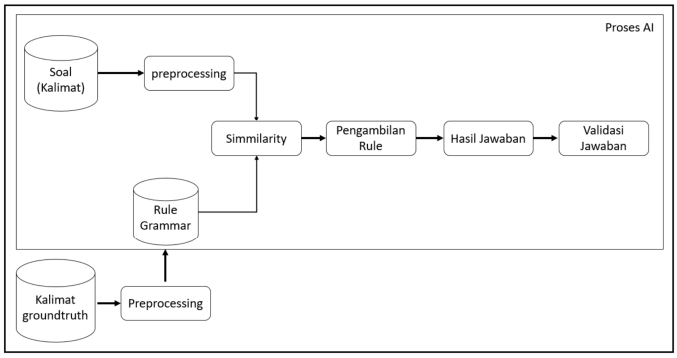


Figure 6. Process learning system in Agent

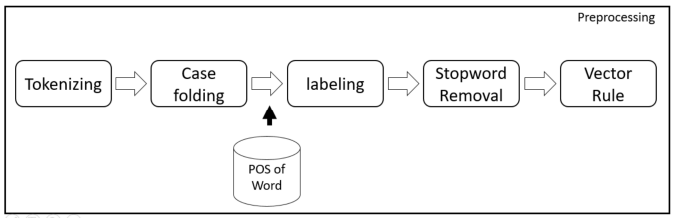


Figure 7. Preprocessing method in Agent

The learning process starts from building database ground truth. This database is consist of POS (part of speech) English words and sentence. After that, agent is converting the sentence into POS and make the rule of grammar using preprocessing method. So in this step the database also has one data called rule grammar. This rule will be used for answer a question in battle menu.

In the battle menu, agent read the English question and answer options. Agent convert the question and answer option into POS English words using preprocessing method. Next, the POS of sentence question will be compared with rule grammar in database. In this process, similarity value for each rule in database will be calculated. the calculation is using SMC and dynamic gram similarity method. The example calculation of dynamic gram is shown in Figure 8.

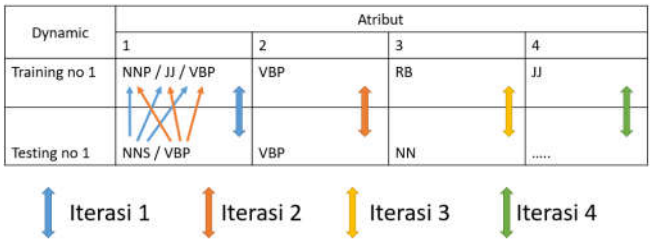


Figure 8. Dynamic gram simmilarity calculation

The highest simmilarity value in rule grammar will be taken for next process. After agent has one rule, it will be compared to answer options. The same POS words in answer option with POS in rule will be chosen. And then the agent can answer the English question in battle scene. The process learning will be start again in the next question until the battle is ended.

III. Result and Discussion

The result of this research, we publish this game on the internet. We also tested system game by our self in conditional environment. The environment consist of smartphone platform android and desktop platform with executable extension. We tested this game for desktop platform in personal computer. The result of system performance assessment in our environment was no lags or good performance. We publish our game into http://www.turtletrainer.blogspot.com. User interface of this game is shown in figure 9, figure 10, figure 11, and figure 12.



Figure 9. Interface of Select Level



Figure 10. Interface of battle menu in game

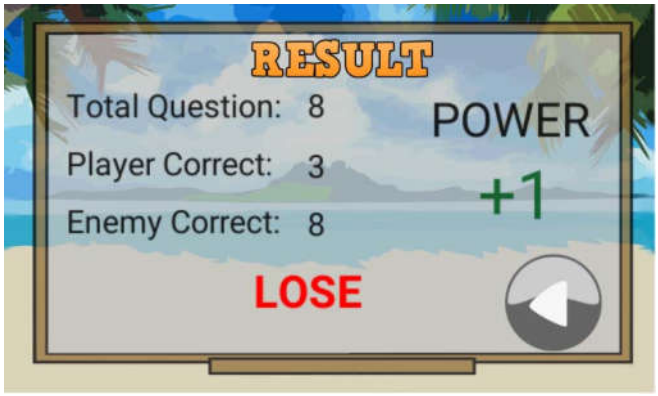


Figure 11. Interface of Result Battle



Figure 12. Interface game in Android device

The evaluation is calculating accuracy of the right answer. The value of accuracy is compared with other method such as unigram, bigram and trigram. The table result of accuracy is showed in Table 1 for gram method and Table 2 for non-gram method.

Table 1. Accuracy gram method

Experiment	Average accuracy gram method			
	Unigram	Bigram	Trigram	Dynamic
Level 1 - 5	60 %	62 %	60%	72 %
Level 6 - 10	52 %	50 %	56%	54 %
Average	57 %	56 %	58 %	66 %

Table 2. Accuracy non gram method

Experiment	Average accuracy non gram method				
	Rand	Always A	Always B	Even A Odd B	Even B Odd A
Level 1 - 5	53 %	48 %	52 %	54 %	46 %
Level 6 - 10	55 %	56 %	44 %	52 %	48 %
Average	54 %	52 %	48 %	53 %	47 %

Gram method is based on n-gram similarity method. Non gram method is classic method in game such as random and consistent options answer. The result indicates that, dynamic gram is better accuracy than unigram, bigram, and trigram. Nevertheless, gram method is better than classic method.

Despite of accuracy, running time evaluation of dynamic method is slower than other method. It can be happened because dynamic gram is checking other possibility of POS English word. The example word is swimming which can be noun or verb-ing because the grammar English has rule called gerund.

But dynamic gram can handle the ambiguous POS English word better than other method. Because we assume, the priority of system performance is accuracy. Furthermore in system game, agent can answer accurately using proposed method. We hope this proposed method can implement other game like action, racing, and battle game which is needed text learning.

IV. Conclusion

From the evaluation, we can conclude that dynamic gram is suitable method for agent to answer automatically in educational game. The accuracy of dynamic gram is 66% which is better than other method. We know this game still has a lot of lack and must improve to make better game especially for agent. So in future research, we will improve and develop this game using deep learning method to answer English Question. Nevertheless we will add gamification element to increase enjoyable for player.

V. References

[1] A. Esmaeili, Z. Ghorrati, and E. Matson, "Multi-agent Cooperation Using Snow-Drift Evolutionary Game Model: Case Study in Foraging Task," in *2018 Second IEEE International Conference on Robotic Computing (IRC)*, 2018, pp. 308–312.

[2] A. Kuz, M. Falco, and R. S. Giandini, "Combining software agents and gaming through student play, an educational module in Agent SocialMetric," in *2017 XLIII Latin American Computer Conference (CLEI)*, 2017, pp. 1–7.

[3] R. E. Dianty, A. M. Azhari, M. F. Al Hakim, I. Kuswardayan, A. Yuniarti, and D. Herumurti, "First aid simulation game with finite state machine model," in *2015 International Conference on Information & Communication Technology and Systems (ICTS)*, 2015, pp. 157–162.

[4] M. Abdi, D. Herumurti, and I. Kuswardayan, "Analisis Perbandingan Kecerdasan Buatan pada Computer Player dalam Mengambil Keputusan pada Game Battle RPG," *JUTI J. Ilm. Teknol. Inf.*, vol. 15, no. 2, p. 226, Jul. 2017.

[5] H. Singal, P. Aggarwal, and V. Dutt, "Modeling Decisions in Games Using Reinforcement Learning," in *2017 International Conference on Machine Learning and Data Science (MLDS)*, 2017, pp. 98–105.

[6] D. Read and Y. Grushka-Cockayne, "The similarity heuristic," *J. Behav. Decis. Mak.*, vol. 24, no. 1, pp. 23–46, Jan. 2011.

[7] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. 2012.

[8] W. H.Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," *Int. J. Comput. Appl.*, vol. 68, no. 13, pp. 13–18, 2013.

[9] C. D. Manning, P. Raghavan, and H. Schütze, "Evaluation in information retrieval," in *Introduction to Information Retrieval*, Cambridge University Press, 2008, pp. 139–161.

[10] D. Jurafsky and J. H. Martin, *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, 2009.

[11] A. A. Yunanto and S. Rochimah, *Jurnal Informatika*, vol. 4, no. 1. Unspecified, 2017.

[12] A. A. Yunanto, D. Herumurti, and I. Kuswardayan, "Kecerdasan Buatan Pada Game Edukasi Untuk Pembelajaran Bahasa Inggris Berbasis Pendekatan Heuristik Similaritas," *J. Sist. dan Inform.*, vol. 11, no. 2, pp. 16–27, May 2017.

A Performance Comparative of Vertical Fragmentation Table using Bond Energy and Graph Based Vertical Partitioning Algorithm

Erba Lutfina[1]
Faculty of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
erbalutfiana@gmail.com

The cost and the speed of data access on a query depends on the numbers of attributes. It is because the query function will search for data from various attributes of the database table, whereas not all the attributes needed on the accessed tables. To separate the attribute that is not necessary, it can be done using the Vertical Fragmentation methods. The purpose of this research aims to compare the vertical fragmentation of database tables using Bond Energy algorithm and Graph-Based Vertical Partitioning algorithm to create a distributed database. The results of both algorithms will determine which attributes will be separated. Then, the results of the two algorithms are compared and evaluated using Partition Evaluator to show the amount of the access data costs that are used, so we can get a conclusion which is better algorithm to be applied. Based on the experimental results, GBVP algorithm produce better partition with fewer computational complexity and generate higher Partition value. 5 Partition Evaluator table has a higher value respectively at 4.98; 10.41; 17.95; 4.64; and 14.07. For the query execution time, 6 tables produces a lower execution time by using GBVP algorithm with the value of each table at 0.003s; 0,002s; 0,002s; 0,002s; 0,003s and 0,003s.

Keywords—*Vertical Fragmentation, Bond Energy Algorithm, Graph-based Vertical Partitioning Algorithm, Distributed Database, Partition Evaluator*

I. INTRODUCTION

A large and complex database system can lead to a decrease in execution time and increased the costs of data access in an information system. A decrease in execution time and an increase in cost happens because access query function will search the data of the various attributes in the database table, whereas not all the attributes needed on the accessed tables. To improve the execution time and reduce the cost of access to data in the database can be resolved through designing a Distributed Database. But in the process of designing a Distributed Database is highly complicated, so the data fragmentation scheme is used to simplify the design process of Distributed Database [1].

Fragmentation scheme divide database tables into fragments that shall be allocated into a site via a network [2].The fragmentation process is divided into two categories: Vertical Fragmentation and Horizontal Fragmentation. Horizontal Fragmentation partitioning relation into disjoint tuple and for each line of the relation will be placed in one or more fragments. Meanwhile in the process of vertical fragmentation, attributes in the relation will be partitioned and placed on a separate fragment [3].

The process of database partitioning will be performed using a vertical fragmentation method based on the calculation results of Bond Energy algorithm (BEA) and Graph-Based Vertical Partitioning algorithm (GBVP) [4], the calculation results will be compared in order to know which algorithm that generates better performance such as execution time and access cost. BEA and GBVP algorithm are used in this study because that two algorithms are the most frequently used in the process of vertical partitioning table.

BEA is an algorithm used in the Vertical Fragmentation process in distributed database[4]. BEA is divided into two algorithms, first algorithm is used to put the group of related data, while the second algorithm is used as the determinant point to create a piece of cluster data. While GBVP is an algorithm that generates fragments using the graph method. GBVP algorithm change affinity matrix into graph affinity for partitioning the fragment with rules and measures that have been defined in GBVP algorithm [5].

Fragmentation results from two algorithms are compared and evaluated using Partition Evaluator (PE) [6] to determine which fragmentation algorithm is more optimal. Partition Evaluator is a function to compare and evaluate different algorithms by totaling the minimum access cost of relevant attributes and the access cost of irrelevant attributes.

The use of GBVP and BEA algorithm is often done by some researchers such as Research using BEA algorithm to

identify the table response time which is vertically partitioned [5], compare the Bond Energy Algorithm with the Modified Bond Energy Algorithm) [7], analyzes the vertical fragmentation table using bond energy algorithm to identify the speed of query execution [8]. In Made's [8] research showed that vertical fragmentation can speed up the query execution time, but it does not apply to the query operations such as update, insert and delete. In that research also analyze the query execution comparison of vertically partitioned tables and the tables that have not fragmented from centralized database. One shortcoming of this comparison is that the result of both query execution time will be very different because the objects and the methods are very different. Therefore, that studies need to be developed again by adding a method that use the same vertical fragmentation, one of them is GBVP algorithm.

This research aims to determine which algorithm is more optimal and more suitable with the dataset to design a distributed database that vertically fragmented using BEA algorithm and GBVP algorithm as well as comparing the access data cost and the query execution time.

The rest of the paper is organized as follows: Section II presents the related work about this research. Section III introduce the research method that used in this research. Section IV discuss our experiment and result. Section V provide the result and conclusion of the research.

II. RELATED WORK

Several Vertical Fragmentation method for Distributed Database were proposed. In Bhaskar et.al [9]research, vertical fragmentation process is performed using Bond Energy Algorithm and apply only one algorithm without collaboration with other algorithm to build a distributed database so the result is less than optimal. The result obtained is the response time decreased but there is no mention of the accuracy details on the paper.

Dariusz and Anna [10] performed a research on the computing complexity of database by using Graphical Partitioning algorithm to perform vertical fragmentation of the database. Just like Bhaskar's research [9], Dariusz and Anna's research also applied only one algorithm without collaboration with other algorithms for optimization process so that the accuracy level is less than optimal. The results showed decreased computational complexity but data access performance still has high access data cost due to parallel processing of subgraph.

Rahimi et,al [4] proposed a modification of the BEA algorithm named Modified Bond Energy Algorithm to improve the vertical fragmentation of a database in the context of performance improvement. The results of this study only minimize the cost of data allocation, but there is still a decrease in access time which may reduce the database performance.

In Made's [8] research showed that vertical fragmentation can speed up the query execution time, but it does not apply to the query operations such as update, insert and delete. In that research also analyze the query execution comparison of vertically partitioned tables and the tables that have not fragmented from centralized database. One shortcoming of this comparison is that the result of both query execution time will be very different because the objects and the methods are very different

III. RESEARCH METHOD

A. Bond Energy Algorithm

Severance, Hoffer and McCormick [10] introduced Bond Energy Algorithm using affinity matrix as an input. Contribution column can be expressed as follows:

$$Cont(Ai, Ak, Aj)=bond(Ai, Ak)+bond(Ak, Aj)- bond(Ai, Aj) \quad (1)$$

Where the bond between attribute i and j are described by the formula

$$bond(Ai, Aj)= \sum_{z=1}^{n} aff(Ai, Ak)\, aff(Ak, Aj) \quad (2)$$

Affinity(aff) calculations of attribute Ai and Ak is based on the access frequency (freq) of query k and access per-execution (acc) of query k, the equation for affinity formula is presented as follows:

$$aff(Ai, Aj)=\sum_{q_k \text{ref } A_i=1} \sum_{q_k \text{ref } A_j=1} freq(q_k) * acc(q_k) \quad (3)$$

Then calculating the split maximize quality value using the following formula:

$$sq = acc(VF1) * acc(VF2) – acc(VF1, VF2)^2 \quad (4)$$

Where acc is a function that calculates the amount of access in the first fragment, the second fragment and combination value of the first and second fragment. Split Quality that has the greatest value will gives the possibility where the fragment will be placed.

B. Graph Based Vertical Partitioning Algorithm

Graph Based Vertical Partitioning Algorithm (GBVP) using Affinity matrix as an input where the affinity between two attributes will be represented with an edge to forms a linearly connected spanning tree [3]. 5 algorithm that need to be considered to produce a vertical fragment using GBVP algorithm is:

1. Construct the Affinity graph of the object attributes.
2. Choosing a randomly selected node.
3. Choose an edge that should be connected to the linear tree that already constructed.
4. If the next edge forms a primitive cycle and the node cycle does not exist, check all the possibilities cycle and if there is a possibility, mark the cycle as the affinity cycle. If the cycle node exist, cut the edge and go to step 3.

5. If the next selected edge does not form a cycle, and no former edge (a selected edge among the last piece and node cycle) and there is a candidate partition, then check the possibility for cycle extension of the new edge. If former edge exists, change the cycle node and check the extension possibility of the cycle by the former edge. If there is no possibility, cut the edge and consider the cycle as a partition. Go to step 3.

C. *Partition Evaluator*

Multiple Linear Regression (MLR) is a regression analysis whose dependent variable (Y) has a relationship with more than one independent variable (x) but the diagram still shows a linear relationship [11].

Partition Evaluator can be used to compare and evaluate different algorithms by adding the value of the irrelevant local attribute access cost and the relevant remote attribute access cost. The equation for irrelevant local attribute access cost is presented as follows:

$$E_M^q = \sum_{i=1}^{M} \sum_{t=1}^{T} q_t^q * |R_{itk}| * \left(1 - \frac{|R_{itk}|}{n_{itk}}\right) \quad (5)$$

Where |R_itk| in this formula is the number of relevant attributes in a fragment. While the equation for Relevant local attribute access cost is presented as follows:

$$E_R^q = \sum_{t=1}^{T} \min \left\{ \sum_{i=1}^{M} q_t^q * |R_{itk}| * \frac{|R_{itk}|}{n_{itk}} \right\} \quad (6)$$

|R_itk| in this formula is the number of relevant attributes in another fragment. Hence the PE is given by:

$$PE = E_M^2 + E_R^2 \quad (7)$$

where,
T is the number of transactions that are under consideration.
qt is the trannsaction frequency of t, for t= 1, 2,….., T.
M is the number of fragments of a partition.
n_itk^r is the number of attributes in the fragments k, which is accessible and associated with fragments i by transaction t.
R_itk is the number of relevant attributes in the fragments k, which is accessible and associated with fragments i by transaction t.

IV. EXPERIMENT AND RESULT

In this research, data were collected from the Academic Information System SMK Texmaco Semarang (SIATEX). The information is about the teacher data, students, value information, attendance and others. There are 10 tables in the respective SIATEX database such as employee, student, score, attendance, study program, subjects, class students, homeroom teacher, schedule, and class. The query used in the vertical table fragmentation amounted to 74 queries.

The data is divided into several stages including query selection, attribute selection and access of each query, affinity calculation between attributes using Bond Energy algorithm and Graph Based Vertical Partitioning algorithm.

The first step is to form n x n matrix affinity as the basis for the table fragmentation process that will be conducted. The next step is calculating the contribution value of attributes using the BEA algorithm by taking two columns randomly from the Affinity matrix. After that, calculate the number of each fragment that is formed, then calculate the maximaze split quality value of each fragment.

GBVP algorithm also uses the same affinity matrix as previously used in the BEA algorithm to execute all steps in the GBVP algorithm.

From the two algorithm process above, two fragments of BEA and GBVP algorithms are produced, which are then calculated and evaluated using Partition Evaluator.

Results of the access data cost calculation using Partition Evaluator show the difference value of BEA and GBVP algorithms as shown in Table 1

Table 1. Access data cost result

No	Table	Partition Evaluator	
		BEA	GBVP
1	ms_pegawai	4,75	4,98
2	tb_absensi	7,41	10,41
3	tb_nilai	17,85	17,95
4	ms_siswa	10,98	8,34
5	ms_prodi	3,49	4,64
6	ms_mapel	4,34	4,34
7	tb_siswakelas	5,32	5,32
8	tb_walikelas	3,72	3,72
9	tb_jadwal	15,28	10,3
10	tb_kelas	12,5	14,07

In the table shown above, 5 of the 10 tables are partitioned using GBVP algorithm has higher Partition Evaluator value than using BEA algorithm, and 3 other tables have same Partition Evaluator value.

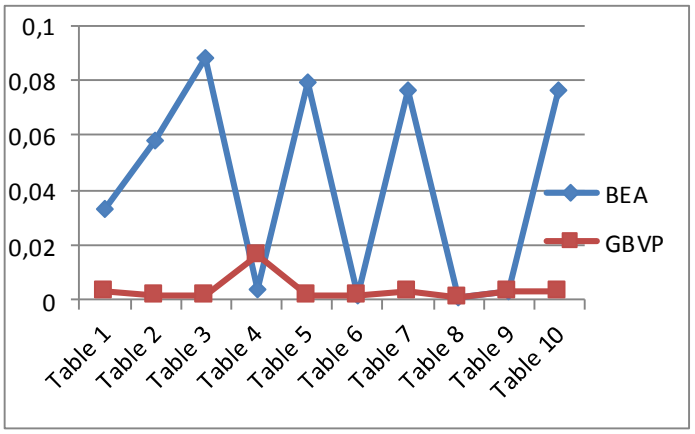


Figure 1. Access Data Cost comparation

From the experimental results show that the GBVP algorithm produces a better fragmentation table than the BEA algorithm, it is seen from the result of Partition Evaluator. The results of the proposed algorithm also show the comparison of query execution time on fragmented tables by using GBVP algorithms and BEA algorithms.

Table 2. Query execution time

	BEA	GBVP
Table 1	0.033 second	0.003 second
Table 2	0.058 second	0.002 second
Table 3	0.088 second	0.002 second
Table 4	0.004 second	0.016 second
Table 5	0.079 second	0.002 second
Table 6	0.002 second	0.002 second
Table 7	0.076 second	0.003 second
Table 8	0.001 second	0.001 second
Table 9	0.003 second	0.003 second
Table 10	0.076 second	0.003 second

Result of the execution time comparison is obtained from implementation of table fragmentation produced by BEA and GBVP algorithm when designing Distributed Databases. Comparison of the execution time is shown in figure 2.

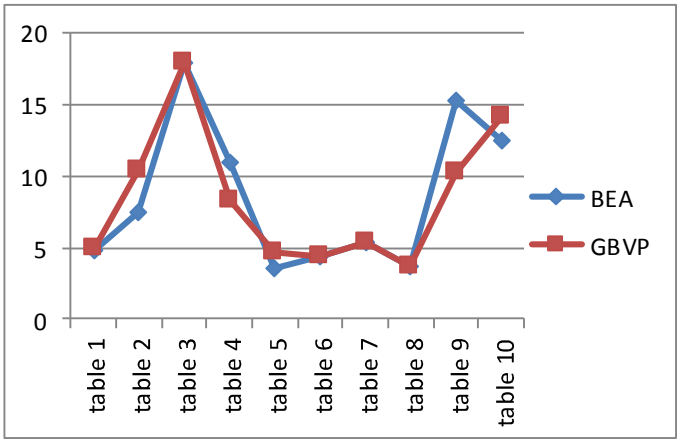


Figure 2. Query execution time comparation

V. CONCLUSION

Based on the results of trials that have been conducted show that the GBVP algorithm have better performance than the previous studies [10] and suitable for designing a distribution database with lower access time values and high partition evaluator values.

GBVP algorithm also produce better partition than BEA algorithm with fewer computational complexity and generate higher Partition value. 5 Partition Evaluator table has a higher value respectively at 4.98; 10.41; 17.95; 4.64; and 14.07. For the query execution time, 6 tables produces a lower execution time by using GBVP algorithm with the value of each table at 0.003; 0,002; 0,002; 0,002; 0,003 and 0,003. For the development and future research can be developed by creating an interactive design tool that allow users to create vertical fragmentation tables and can be used to find out the information of the data access cost.

REFERENCES

[1] D. C. Gope, "Dynamic Data Allocation Methods in Distributed Database System," vol. 4, no. 6, 2012.

[2] R. Journal and O. F. Information, "Clustering methods in data fragmentation 1," vol. 14, no. 1, pp. 81–97, 2011.

[3] C. Science, O. Technology, and T. Nadu, "Vertical Fragmentation , Allocation and Re-Fragmentation in Distributed Object Relational Database Systems- (Update Queries Included) Surabhi Gupta , 2 Shruti Panda," vol. 4, no. 7, pp. 45–52, 2012.

[4] H. Rahimi, F. Parand, and D. Riahi, "Hierarchical simultaneous vertical fragmentation and allocation using modified Bond Energy Algorithm in distributed databases," *Appl. Comput. INFORMATICS*, 2015.

[5] R. M. H. Al-sayyed, F. A. Al Zaghoul, D. Suleiman, and M. Itriq, "A New Approach for Database Fragmentation and Allocation to Improve the Distributed Database Management System Performance," no. October, pp. 891–905, 2014.

[6] A. Bihani, M. S. Reddy, and A. Chandrasekharan, "A new approach to re-fragment in a distributed database environment," *Indian J. Sci. Technol.*, vol. 9, no. 3, pp. 1–8, 2016.

[7] P. Koszali, *Zeszyty Naukowe Wydziału Elektroniki i Informatyki*, no. 6. 2014.

[8] F. Matematika, P. Alam, and U. Udayana, "Analisis Fragmentasi Tabel secara Vertikal Menggunakan Algoritma Bond Energy dalam Hubungannya dengan Kecepatan Eksekusi Query, Made Hanindia Prami Swari 1 , Ngurah Agus Sanjaya ER 2 Program Studi Teknik Informatika , Jurusan Ilmu Komputer ," vol. 5, no. 1, pp. 11–18, 2012.

[9] R. Bhaskar, "An Analysis of Vertical Splitting Algorithm," vol. 52, no. 18, pp. 30–36, 2012.

[10] A. Bernatowicz, "Application of correlation in the vertical fragmentation based on statistic of queries."

[11] V. Galaktionov, "for Vertical Partitioning Problem : an Initial Performance Study," pp. 24–31.

QoS Improvement Analysis of VoIP Service Which Uses Overlay Network. Case Study: Calling AWS VoIP Gateway From Bandung, Indonesia

Indrajati Haryono[1], Dodi Wisaksono Sudiharto [2], Aji Gautama Putrada[3]
School of Computing
Telkom University
Bandung, Indonesia
indrajatiharyono@gmail.com[1], dodiws@telkomuniversity.ac.id[2], ajigps@telkomuniversity.ac.id[3]

***Abstract*—VoIP (Voice over Internet Protocol) has been widely adopted by both a public and a private entities, making this service as the one of popular services over time. However, up to now, how to maintain quality of VoIP service which is according to the standard quality has always become a central issue for studying. Especially, if the service of a public computer network which is available in a region or a country is still lack. For example, in terms of adequate bandwidth limitation. To overcome this condition, a required mechanism which can improve the quality of VoIP service is needed, so the user can be satisfied in using this service. This study is going to explore the mechanism of VoIP service quality improvement which is done by using overlay network. The overlay network is applied by making a tunnel on the network. In addition, VoIP gateway is placed on cloud system which is Amazon Web Service (AWS). Therefore, the placement of the gateway is going to resolve the hardware gap or to maintain the computing-related barrier. The measurement of this service is done by making a call service from Bandung, Indonesia to VoIP gateway which is placed on AWS cloud system.**

Keywords—VoIP; overlay network; tunneling; cloud system.

I. INTRODUCTION

In general, voice service run on the POTS (Plain Old Telephone Service). Nowadays, The development of technology makes voice service can also run on IP network. Its popular service name is VoIP. The use of VoIP has been widely applied in some applications on the market. However, not all of existing VoIP services have the same system. Especially, if the service which is implemented has a mechanism to improve the quality of voice service, such as by using overlay network.

By using traditional VoIP technology, the voice is transmitted over a network through multiple servers. Each server may not be located in the same area with others. This condition can increase a communication maintenance cost, and also can reduce the quality of VoIP service. In addition, common problems for maintaining VoIP network can also rise [1]. On the other hand, the presence of delay and packet loss becomes the problems for maintaining VoIP service [2]. In some cases, the best solution for maintaining VoIP service infrastructure is by using cloud system [3]. Then, this study is going to make the VoIP system by adopting the cloud environment.

According to the delay and the packet loss, there is a method which can be used to minimize them. The method is overlay network. The overlay network method which can be used such as socket-based tunneling method, and there is also a routing daemon for packet delivery [4].

The study focuses on the mechanism of network overlay by using tunneling to improve the quality of VoIP service. Parameters measured are delay and packet loss. The calculation of Mean Opinion Score (MOS) is used to determine the quality of VoIP service [5].

II. RELATED WORKS

There are some references related to the VoIP service issue. In [6], tunneling mechanism is used to improve the quality of VoIP service. This can be generated by prioritizing certain protocols which pass through the tunnel.

In other study, it is mentioned that cloud-based VoIP system can simplify an infrastructure operation [3]. The ability of the infrastructure to flexibly change can create the VoIP system environment to be changed as needed.

According to the literatures which are mentioned, this study is going to design the VoIP system using tunneling mechanism, and placing the VoIP gateway on cloud service provider which is the AWS.

A. VoIP

To transmit a message over the internet using VoIP service, the first step is that a voice analog signal is going to be encoded and converted into frames. Then each frame is encapsulated into a packet, and it is going to be sent to the destination using the protocol. In general, the protocol used is UDP. After that, the receiver is going to decode the frames, and convert them back into the voice analog signal [7].

Related to the packet delivery on the VoIP service, there are 2 elements which are important, namely a protocol and a codec. The VoIP protocol determines how voice packets are transmitted over the network. Meanwhile, the codec role is at

the reception which is going to use a buffer mechanism to overcome delay packets, and is going to use Forward Error Correction (FEC) mechanism, or Packet Loss Concealment (PLC) mechanism to handle the impact of packet loss, or delay which is exceed [7].

1) SIP

This study is going to apply a popular protocol which is standardized by the IETF (Internet Engineering Task Force), the SIP (Session Initiation Protocol) [8]. The SIP uses an URI (Uniform Resource Locator) to identify the logical destination. So that, it does not use IP address for addressing. The information used for addressing can be: name, e-mail address, or phone number [5].

2) G.711

Besides the required protocol, VoIP system also requires a codec to change the analog signal to the digital signal. This codec uses PCM (Pulse Code Modulation) for sampling an 8 bit sound signal every 0.125 ms [9]. This study uses G.711 codec because it is the first codec for VoIP system and has proven to be a popular codec.

The followed graph is about the type of codec faces MOS (Mean Opinion Score) [10].

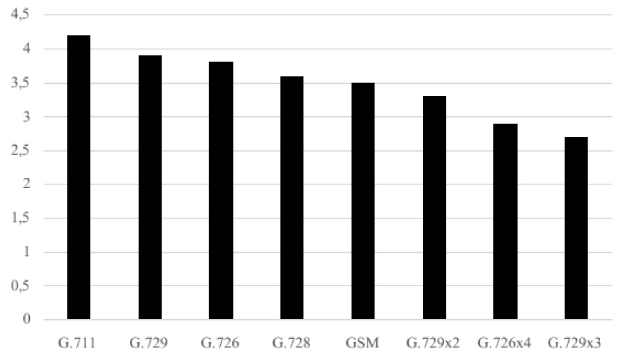


Fig. 1. Grafik codec terhadap MOS

3) Amazon Elastic Compute Cloud (EC2)

The AWS has many services which are grouped into 5 major parts which are compute service, storage service, database service, analytic service, application service, and deployment service [11]. This study uses Asterisk as a VoIP gateway, so that the use of the compute service is needed, and it is provided by the Amazon Elastic Compute Cloud (EC2).

The Elastic Compute Cloud (EC2) has two access control mechanisms. One of them is that the EC2 has a tool which can write SSH (Secure Shell) public key to AMI (Amazon Machine Image), which allows user who has private key to login to a superuser account. The AWS also has a firewall which can be used to restrict a network connectivity based on IP address and TCP port number [2]. Those mechanisms are very useful as a security feature when there is VoIP call.

The Amazon EC2 has several server regions which can be chosen for the usage. They can be seen on the followed table [12]:

TABLE I. THE LIST OF AMAZON REGIONS.

Code	Name
us-east-1	US East (N. Virginia)
us-east-2	US East (Ohio)
us-west-1	US West (N. California)
us-west-2	US West (Oregon)
ca-central-1	Canada (Central)
eu-west-1	EU (Ireland)
eu-central-1	EU (Frankfurt)
eu-west-2	EU (London)
ap-northeast-1	Asia Pacific (Tokyo)
ap-northeast-2	Asia Pacific (Seoul)
ap-southeast-1	Asia Pacific (Singapore)
ap-southeast-2	Asia Pacific (Sydney)
ap-south-1	Asia Pacific (Mumbai)
sa-east-1	South America (São Paulo)

There are many choices of computation which are provided by Amazon EC2. They can be seen on the followed table:

TABLE II. THE LIST OFAWS COMPUTATION TYPE..

Instance Family	Type
General purpose	t2.nano | t2.micro | t2.small | t2.medium | t2.large | t2.xlarge | t2.2xlarge | m4.large | m4.xlarge | m4.2xlarge | m4.4xlarge | m4.10xlarge | m4.16xlarge | m3.medium | m3.large | m3.xlarge| m3.2xlarge
Compute optimized	c4.large | c4.xlarge | c4.2xlarge | c4.4xlarge | c4.8xlarge | c3.large | c3.xlarge | c3.2xlarge | c3.4xlarge | c3.8xlarge
Memory optimized	r3.large | r3.xlarge | r3.2xlarge | r3.4xlarge | r3.8xlarge | r4.large | r4.xlarge | r4.2xlarge | r4.4xlarge | r4.8xlarge | r4.16xlarge | x1.16xlarge | x1.32xlarge
Storage optimized	d2.xlarge | d2.2xlarge | d2.4xlarge | d2.8xlarge | i2.xlarge | i2.2xlarge | i2.4xlarge | i2.8xlarge | i3.large | i3.xlarge | i3.2xlarge | i3.4xlarge | i3.8xlarge | i3.16xlarge
Accelerated computing	f1.2xlarge | f1.16xlarge | p2.xlarge | p2.8xlarge | p2.16xlarge | g2.2xlarge | g2.8xlarge | g3.4xlarge | g3.8xlarge | g3.16xlarge

In this study, the AWS region used is located in Oregon, USA. The computation selected is t2.micro which is free payment for 12 months and it is supported by Amazon Free Tier [13] .

4) Asterisk

The Asterisk is chosen as the VoIP gateway which is using the SIP protocol, and the protocol is used to run the G.711 codec [10]. This choice is selected because of there are many features on Amazon Web Service which support the Asterisk, such as documentations.

5) Softphone

Softphone is a software which is used to make a phone call over internet. It can be used on computer, smartphone, or tablet []. To communicate, minimal, two softphones must be equipped with both supported protocol and voice codec. The softphone usually has many codecs which can be used. In general, the softphone has G.711 as standard protocols. This study is going to use X-Lite softphone on the PC, and is going to apply Zoiper on Android smartphone which can be downloaded for free on the internet.

Differs from the X-Lite, the Zoiper can display a statistic report when it is used for making a call. So, it can be operated to measure the quality of VoIP service.

B. QoS (Quality of Service)

The QoS (Quality of Service) is the main focus in this study, which is how to get delay and packet loss as minimum as possible. The quality of VoIP service means how to be able to listen and to speak clearly and continuously without have any jitter, delay, or packet loss [14].

1) Delay

Delay is a time required by a data packet during transmission to reach a receiver [5]. Delay on a network makes the data packet becomes late to reach the destination. For the usage of VoIP service, the recommended delay is 150 ms [5].

Delay used in this study is one way delay which has a component coder delay, a packetization delay, a serialization delay, and a network delay [15]. The explanation for those components can be followed as:

1. Coder delay is a time required by digital signal processor (DSP) to compress block from PCM (Pulse-Code Modulation) [16].
2. Packetization delay is a time required to fill a packet with an encoded sound [16].
3. Serialization delay is a time required to place a data frame to a network interface [16].
4. Network delay is a time required for the packet to arrive at the destination. In this research, network delay is obtained by using Wireshark.

a) Packet Loss

Packet loss is the percentage of packet loss during a transmission [5]. The amount of packet loss can be calculated by the followed formula:

$$Packet\ loss = \frac{P_d}{P_s} \times 100\% \quad (1)$$

which: P_d= the number of the packet loss.
P_s= the number of the transmitted packet.

b) E-Model

It is an objective measurement which is standardized by ITU-T G.107 for telecommunication network. The output of E-Model is R-Factor which is used to calculate the Mean Opinion Score (MOS) which is the level of the network quality [5].

$$R = 94.2 - I_d - I_{ef} \quad (2)$$

which: I_d = is a factor of the decreased sound quality by the presence of delay.
I_{ef}= is a factor of quality degradation due to compression technique and packet-loss.
R = is a factor of transmitted quality.

which: d = delay.
e = the percentage of packet-loss.
H = heavyside function, which:
H = 0, if x< 0, and;
H = 1, if$x\geq 0$

The I_d and I_{ef} can be calculated by using the followed formula:

$$I_d = 0.024d + 0.11(d - 117.3)H_{(d-177.3)} \quad (3)$$

$$I_{ef} = 7 + 30\ln(1 + 15e) \quad (4)$$

To convert R valueinto MOS can use followed formula [5]:

$R < 0$; MOS= 1, (5)
$R > 100$; MOS= 4.5, (6)
$0 < R < 100$; MOS= $1 + 0.035\ R + (7 \times 10^{-6})\ R\ (R - 60)(100 - R)$ (7)

TABLE III. THE RELATION BETWEEN R AND MOS.

R	MOS	User Satisfaction
90	4.34	Very satisfied
80	4.03	Satisfied
70	3.60	Some users dissatisfied
60	3.10	Many users dissatisfied
50	2.58	Nearly all users dissatisfied

2) Mean Opinion Score

Mean Opinion Score (MOS) is an assessment system of a sound sample which produces a quality score. Here is a table of MOS quality value which is using the value in the range 1 to 5. The quality of voice call through codecs is measured by subjective testing under specified condition to determine the MOS value [10].

TABLE IV. THE SCALE OF MOS VALUE

MOS Value	Quality
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

C. Overlay Network

Overlay network using a tunneling is the main point in this study. Overlay network is a computer network which is created on the top of other network. The overlay network can handle traffic from multiple connections which can be from multiple directions. By using this kind virtual network, each connection can reduce its hop.

1) VPN Tunneling

Tunneling is the principle of creating a tunnel on a public network, so it seems there is a point-to-point relationship with encapsulated data. This tunneling is one of the network overlay mechanisms which can be used [4]. VoIP tunneling provides the way to ensure the confidentiality and the integrity of a call based on Internet Protocol (IP) by encapsulating the data from one protocol to other protocol [17]. This study uses SSL/TLS as the protocol for tunneling. This protocol supports data encryption and data authentication for HTTP traffic. In addition, it can be used to secure RTP (Real-time Transport Protocol) traffic [18]. The application which uses this protocol is OpenVPN which is an open source application.

D. Traffic Geneator

The purpose of traffic generator usage in this study is to emulate the packets on the network. The application which is going to be used as a traffic generator is Ostinato [19]. This application can emulate the traffic which is needed. By using this application, this study is going to make 100 Mbps traffic emulation.

III. PROBLEM DEFINITION

Delay and packet loss can reduce the quality of VoIP service on the transmission of voice data. According to delay and packet loss, there is a method which can be used to minimize it, such as overlay network by making a tunnel based on socket network [4].

In this study, a comparison between VoIP service which uses overlay network (i.e. tunneling) and VoIP service without overlay network is performed. The parameters measured are delay and packet loss.

IV. PROPOSED SYSTEM

A. Hardware and Software Specification

This study utilizes the cloud environment which is provided by AWS. Thus user can specify hardware specification freely. The hardware also can be upgraded and downgraded any time.

The hardware specification which are applied can be followed as:

- The server which is run on AWS has specification:
  - OS : Ubuntu 14.04 LTS
  - CPU : t2.micro
  - RAM : 1 GB
  - HDD : 20 GB
- The PC which has the VoIP softphone installed on it has spec:
  - OS : Windows 10/ Ubuntu 14.04 LTS
  - CPU : Intel Core i7 CPU @2.4 Ghz
  - RAM : 4 GB
  - HDD : 500 GB
- The LG Magna smartphone has specification:
  - CPU : Quad-core 1.2 GHz
  - RAM : 8 GB
  - OS : Android 5.0.1 Lollipop

The software which are used in this study can be followed as:

- Ubuntu 14.04 becomes a server system operation which is used in this study. The client applies Windows to run X-Lite or Zoiper as a softphone.
- Asterisk becomes VoIP gateway, and several SIP accounts are reserved on it which is used for calling.
- OpenVPN Access Server and OpenVPN Client. OpenVPN Access Server and Asterisk are installed as a bundle on Amazon EC2.So that the user only has one instance. On the other hand, OpenVPN Client is applied on the client system which is connected to Open VPN Access Server.
- Wireshark is used to capture delay and packet loss of packets which occur during VoIP call.
- X-Lite as a softphone which is applied to register SIP account on Asterisk, and makes a call between PC clients..
- Zoiper is a softphone which is implemented to register SIP account on Asterisk, and executes a call between Android clients.
- Ostinato is a traffic generator which is applied to increase the load traffic on the network. Ostinato is going to be paired into PCs which is connected each other on a local network.

B. The Topology of VoIP Network

The study is going to compare between VoIP service which does not use overlay network and VoIP service which uses tunneling. The topology network of VoIP service which does not use tunneling network can be seen on the followed picture:

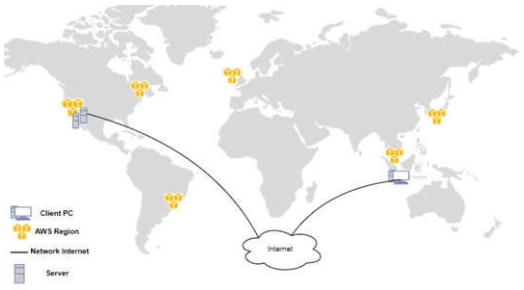


Fig. 2. The topologi network of VoIP service without tunneling

The picture above describes AWS Region which is located in Oregon, USA. By using AWS Region in other part of the world, far from Indonesia, it is hoped that the traffic network is going to pass the number of hops when connected to a PC located in Indonesia. So the environment under the study is going to actually objectively describe the real network condition.

To improve VoIP service, the mechanism of overlay network by using tunneling is executed. The placement of VoIP gateway and PC is followed Figure 2, but the tunneling mechanism is used for improving the VoIP service.

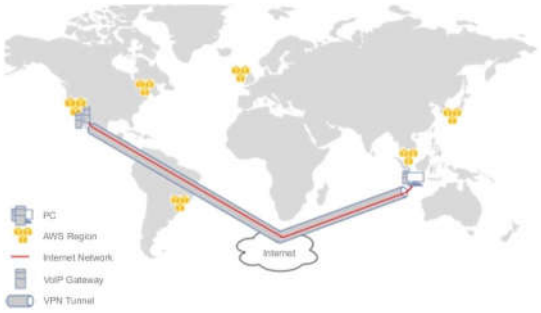


Fig. 3. The topologi network of VoIP service using tunneling

Network traffic which is created by using traffic generator is emulated to represent the presence of dense network on the local area network. This is done because this kind of traffic can be generated by using traffic generator like Ostinato. The Ostinato placement on the local area network can be seen on the followed picture:

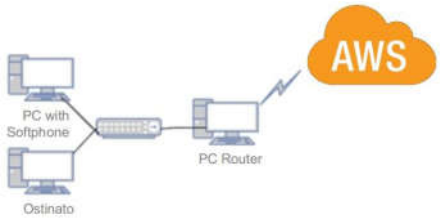


Fig. 4. Penempatan trafik generator pada jaringan lokal

The configuration of the Ostinato uses a burst value of 100 with packets per burst of 10,000. The packet which is sent is 100 Mbps.

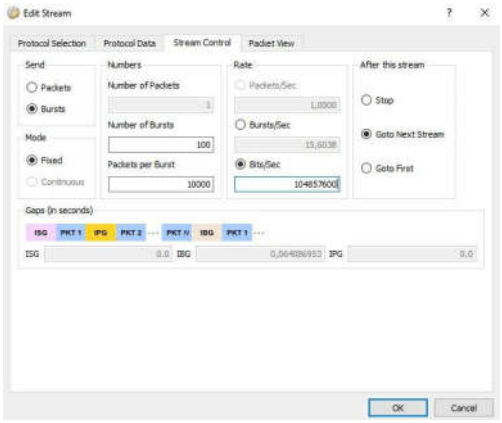


Fig. 5. The Ostinato configuration

Thus, it is hoped that the local area network is going to be full of dense traffic which is created by the traffic generator.

VoIP service runs using topology which is described on Figure 2 is going to be compared with VoIP service which applies topology drew on Figure 3. The measurement results are the value of delay and packet loss, which then are going to be converted into MOS value. By using the MOS value obtained is expected the use of VoIP service using overlay network can increase the quality of the VoIP service which is not using tunneling.

V. ANALYSIS AND DISCUSSION

A. *The Delay and Packet Loss Measurements of VoIP Service Without Overlay Network*

The delay measurement of topology without tunneling is done using Wireshark. The delay value can be captured from the average value in the "time delta from previous displayed frame" column which is presented by Figure 6. The average delay is 20.18 ms.

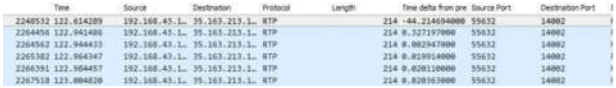


Fig. 6. The delay measurement of VoIP data without overlay network.

On the other hand, the packet loss measurement obtained gives the missing package prosentage of 4.3%.

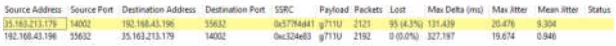
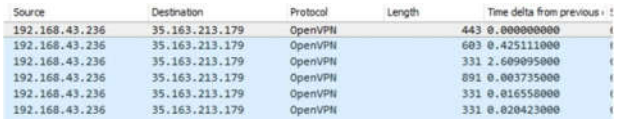


Fig. 7. The packet loss measurement of VoIP data without overlay network.

B. *The Delay and Packet Loss Measurements of VoIP ServiceUsing Overlay Network*

By using Wireshark, it is known that the average value in the "time delta from previous displayed frame" column which is for delay calculation is 21.31 ms.



Source	Destination	Protocol	Length	Time delta from previous
192.168.43.236	35.163.213.179	OpenVPN	443	0.000000000
192.168.43.236	35.163.213.179	OpenVPN	603	0.425111000
192.168.43.236	35.163.213.179	OpenVPN	331	2.609095000
192.168.43.236	35.163.213.179	OpenVPN	891	0.003735000
192.168.43.236	35.163.213.179	OpenVPN	331	0.016558000
192.168.43.236	35.163.213.179	OpenVPN	331	0.020423000

Fig. 8. The packet loss measurement of VoIP service which uses overlay network.

Thepacket lossmeasurement which uses ping, and the procentage of packet loss is 0%.

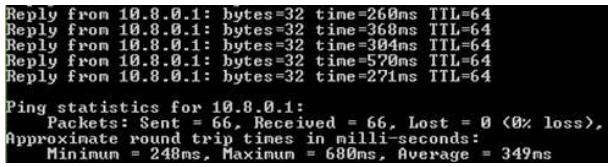


Fig. 9. The packet loss measurement of VoIP servicewhich uses overlay network.

The packet loss measurement of VoIP service which uses overlay network is not applied by using Wireshark because of the limitation of the tool to read the protocol which is used by OpenVPN.

C. The Measurement of MOS (Mean Opinion Score)

The measurement and the analysis of MOS value is found objectively and subjectively. For the objective calculation, the delay and packet loss data obtained in the previous measurement is used. Yet, the subjective measurement can be collected by executing the questionnaire.

1) The Objective MOS Measurement for The Topology Without Tunelling.

The measurement using E-Model can be done by using formula (2),(3), and (4).

The variable of d can be obtained by the accumulation of:

- Coder delay = 0.125 ms
- Packetization delay (G.711) = 20 ms
- Serilization delay (4G Telkomsel) = 1 ms
- Network delay (without tunneling) = 20.18 ms
- d which is obtained = 41.305

So that I_d which is calculated is 0.9913.

The variable of e which calculated for packet loss is 4.3 %. So that I_{ef} which is calculated is 21.932.

R factor which is obtained is 71.2767, by using formula (7). Thus the MOS value is 3.656, for the VoIP service without tunneling.

2) The Objective MOS Measurement for The Topology Which Uses Tunelling.

By using formula (2),(3), and (4), the accumulation of:

- Coder delay = 0.125 ms
- Packetization delay (G.711) = 20 ms
- Serilization delay (4G Telkomsel) = 1 ms
- Network delay (using tunneling) = 21.31 ms
- d which is obtained = 42.435

So that I_d which is calculated is 1.018.

The variable of e for packet loss is 0 %. So that I_{ef} which is obtained is 7.

Then R factor value is 86.092, by using formula (7). Thus the MOS value is 4.231, for the VoIP service which uses tunneling.

Based on both MOS values, it can be said that there is an quality improvement while VoIP service uses tunneling. It clearly shows by the followed table:

TABLE V. THE OBJECTIVE MOS VALUE OF VOIP SERVICE

The Objective MOS Value of VoIP Service	
Without Tunneling	Using Tunneling
3.656	4.231

It can be seen that the VoIP service uses tunneling have good quality compare with the VoIP service without tunneling which have fair quality.

3) The Subjective MOS Measurement

The measurement is done by giving some questions to 30 users. Here are the sampling questions asked:

- Are you satisfied with the quality of VoIP service?
- While making a call, is the voice of the other person sounds clearly?

By using those such questions, judgments are made to get the subjective MOS value. The followed tables are the result of the subjective MOS measurement.

TABLE VI. THE SUBJECTIVE MOS VALUE OF VOIP WITHOUT TUNNELING

User	MOS	User	MOS
1	3	16	4
2	4	17	2
3	3	18	3
4	3	19	4
5	3	20	3
6	2	21	5
7	5	22	4
8	4	23	4
9	4	24	4
10	4	25	4
11	3	26	4
12	3	27	3
13	3	28	5
14	3	29	4
15	3	30	4
Average		3.5	

TABLE VII.
THE SUBJECTIVE MOS VALUE OF VOIP USES TUNNELING

User	MOS	User	MOS
1	3	16	4
2	4	17	3
3	3	18	4
4	5	19	4
5	4	20	3
6	2	21	5
7	5	22	5
8	5	23	4
9	5	24	4

User	MOS	User	MOS
10	5	25	4
11	4	26	4
12	4	27	4
13	3	28	5
14	4	29	4
15	3	30	4
Average		4	

The subjective MOS value between VoIP service without tunneling and VoIP service which uses tunneling refers to Table VII. It shows that according to the quality of VoIP service mentioned on Table IV, the result is the same as the objective MOS assessment. VoIP service without tunneling is can be considered as fair. On the other hand, VoIP service which uses tunneling has good quality.

TABLE VIII. THE SUBJECTIVE MOS VALUE OF VOIP SERVICE

The Subjective MOS Value of VoIP Service	
Without Tunneling	Using Tunneling
3.5	4

VI. CONCLUSION

The measurement of VoIP services without tunneling overlay network capture the delay of 20.18 ms and the packet loss of 4.3 %. On the other hand, the measurement of VoIP service which uses tunneling overlay network get the delay of 21.31 ms and the packet loss equal to 0 %. This happens because there is packet header addition of the VoIP service by using tunneling overlay network. However, the packet loss reduction leads the quality of VoIP service which uses tunneling overlay network. This can be seen in the objective MOS value which increases from the original quality level which is at a fair level for VoIP service without tunneling overlay network, to be at a good level for VoIP service which uses tunneling overlay network. The subjective MOS value measurement also gives exactly the same pattern as the objective MOS value measurement. Based on the information of this study, it can be said that getting AWS VoIP service by make a call from Bandung, Indonesia can have good quality.

REFERENCES

[1] S. A. Patinge and P. D. Soni, "Metamorphosis in VoIP cloud computing services used in VoIP," Int. J. Appl. or Innov. Eng. Manag., vol. 2, no. 2, pp. 236–239, 2013.

[2] J. H. James, B. Chen, and L. Garrison, "Implementing VoIP: A voice transmission performance progress report," IEEE Commun. Mag., vol. 42, no. 7, pp. 36–41, 2004.

[3] F. Gerea, "Implementation of cloud computing into VoIP," Database Syst. J., vol. 3, no. 2, pp. 3–12, 2012.

[4] D. W. Sudiharto, F. A. Yulianto, and A. N. Arista, "Comparative analysis of voice over internet protocol (VoIP) quality on priority queue (PQ) and class-based queue (CBQ) management system using link-sharing mechanism setting," in Int. Conf. Inf. Commun. Technol. (ICoICT), 2015, pp. 419–424.

[5] T. Yildirim and P. J. Radcliffe, "VoIP traffic classification in IPSec tunnels," in Int. Conf. Electron. Inf. Eng. (ICEIE), 2010, pp. 151–157.

[6] Y. Amir, C. Danilov, S. Goose, D. Hedqvist, and A. Terzis, "An overlay architecture for high quality VoIP streams," IEEE Transactions on Multimedia, vol. 8, no. 6, pp. 1250–1262, 2006.

[7] I. Dalgic and H. Fang, "Comparison of H. 323 and SIP for IP telephony signaling," Multimedia Systems and Applications II, vol. 3845, pp. 106-123, 1999.

[8] A. P. Markopoulou, F. A. Tobagi, and M. J. Karam, "Assessment of VoIP quality over internet backbones," in Proceedings.Twenty-First Annu. Jt. Conf. IEEE Comput. Commun. Soc., vol. 1, pp. 150–159, 2002.

[9] B. Goode, "Voice over internet protocol (VoIP)," in Proc. IEEE, vol. 90, no. 9, pp. 1495–1517, 2002.

[10] "What is Cloud Computing?," *AWS*, 2018. [Online]. Available: https://aws.amazon.com/what-is-cloud-computing/. [Accessed: 25-Apr-2018].

[11] "Regions and Availability Zones," AWS Documentation, 2018. [Online]. Available: https://docs.aws.amazon.com/AWSEC2/latest/UserGuide/using-regions-availability-zones.html#concepts-available-regions. [Accessed: 25-Apr-2018].

[12] "Using the Free Tier," AWS Documentation, 2018. [Online]. Available: https://docs.aws.amazon.com/awsaccountbilling/latest/aboutv2/billing-free-tier.html. [Accessed: 25-Apr-2018].

[13] P. Radmand and A. Talevski, "Impact of encryption on QoS in VoIP," in Social Computing (SocialCom), 2010.

[14] M. Voznak and M. Halas, "Delay variation model with RTP flows behaviour in accordance with M/D/1 Kendall's notation," Advances in Electrical and Electronic Engineering, vol. 8, no. 5, pp. 124-129, 2011.

[15] "Understanding Delay in Packet Voice Networks," *Cisco*, 2006. [Online]. Available: https://www.cisco.com/c/en/us/support/docs/voice/voice-quality/5125-delay-details.html#sourceofdelay. [Accessed: 25-Apr-2018].

[16] J. Touch, "Dynamic internet overlay deployment and management using the X-Bone," Comput. Networks, vol. 36, no. 2–3, pp. 117–135, 2001.

[17] E. B. Fernandez, J. C. Pelaez, and M. M. Larrondo-Petrie, "Security patterns for voice over IP networks," in Int. Conf. Comput. Glob. Inf. Technol. (ICCGI), p. 33, 2007.

[18] W. B. Diab, S. Tohme, and C. Bassil, "VPN analysis and new perspective for securing voice over VPN networks," in Int. Conf. Netw. Serv. (ICNS), pp. 73–78, 2008.

[19] "Network Traffic Generator and Analyzer," Ostinato, 2018. [Online]. Available: https://ostinato.org/. [Accessed: 25-Apr-2018].

Surveillance Embedded IP Camera with Integrated Cloud Storage and Image Deblurring Using Richardson-Lucy

Almer Sofyan[1], Dodi Wisaksono Sudiharto[2], Catur Wirawan Wijiutomo[3]
School of Computing
Telkom University
Bandung, Indonesia
almersofyan@gmail.com[1], dodiws@telkomuniversity.ac.id[2], caturwijiutomo@telkomuniversity.ac.id[3]

***Abstract*—IP camera has been widely used to monitor an area for security purpose. Generally, this monitoring is done to supervise certain area which requires high privilege, so the IP camera can record all events which are occurred. Recording can be used to track what happened as if there was a law violation in the supervised area. However, when image data is retrieved, sometimes can occur damage due to a monitored object is moving. The image which has been damaged by the movement of this object is termed blur. Thus, to fix it is required image processing method such as deblurring. An image data also needs to be saved in the storage. By using cloud storage, it is going to get a large storage space and can also eliminate additional cost such as the maintenance of server storage. In this study is designed and is implemented the prototype of Surveillance Embedded IP Camera which has image processing feature (deblurring) using Richardson-Lucy method and the integration of storage using the cloud.**

Keywords—Surveillance Embedded IP Camera; Richardson-Lucy; deblurring; cloud storage.

I. INTRODUCTION

Although many entities have IP cameras for monitoring purpose, the recording may still manual [2]. In fact, there is typical IP camera which does not have a recording feature which is also connected to the internet [3]. As a result, when there is unlawful action in an area, it is difficult to trace quickly.

In general, some IP cameras have been able to record data which are going to be stored in a storage device [3]. Nonetheless, the storage which is used may still such as a memory card like SD card or flash memory [2]. It needs to be done manually if the data are going to be moved to other device such as PC (Personal Computer).

In addition, when IP camera performs passive surveillance such as taking data over a period of time, sometimes there is data damage in data retrieval due to moving object when the recording event is done. This damaged image is termed as blur [6]. The blur which is caused by moving object when the IP camera shoots is called as motion blur [9] [14]. To correct motion blur needs the image processing. The image processing method that can be done on motion blur is deblurring. There are several techniques that can be used for deblurring such as Van Cittert, Landweber, Richardson-Lucy, Poisson Map, and Laplacian Sharpening Filter [11].

This study uses an improvement on the motion blurred image using Richardson-Lucy algorithm. The algorithm is chosen for the reason that a high PSNR (Peak Signal to Noise Ratio) value can be generated without being affected by noise on the image [11].

II. RELATED WORKS

There are some references related to the surveillance camera issue such as:

- In [6] describes that object which is moving during recorded by IP camera can generate a blurred image.
- In [9] presents that blurred image which is generated by IP camera can be reduced using Richardson-Lucy algorithm.
- In [11] explains that the Richardson-Lucy algorithm has high PSNR value result and it is not much affected by the image noises.
- In [4] shows that the use of IP camera which supported by cloud storage becomes cost efficient.
- In [5], establishes an explanation that cloud, whether it is a saving media or a processing, its installation process can be difficult.
- In [15] exhibits that Raspberry Pi with ARM9 processor is a technology which is cheap, easy to obtain, and commercially available, so it can be classified as a technology that has cost-effective element.

Based on those literature studies, the design and the implementation of Surveillance Embedded IP Camera which is going to be generated has several features such as:

- The ability to present a passive surveillance [15].
- It can do an image processing (deblurring) using Richardson-Lucy algorithm [6] [9], and;
- Its storage is integrated with cloud storage [4].

III. PROBLEM DEFINITION

Blur is a typical damage which happens to the image and it makes the image look unclear (blur). This can be occurred due to camera movement during shooting or as a result of moving object when its picture is taken [6]. The typical damage is known as motion blur [9] [14].

To restore motion blur, a deblurring mechanism can be performed. The deblurring can be done using a variety of algorithms. However, this study uses is Richardson-Lucy algorithm. This selection is done by the reason that processing the image is not affected by the type of noises which are generated after image processing is done [6]. Although the execution time of Van Cittert algorithm and Landweber algorithm are faster than Richardson-Lucy algorithm, they are sensitive to noises [6] [17]. Differs from the Richardson-Lucy algorithm which is not affected by the type of noises created after image processing is executed [6] [17].

IV. PROPOSED SYSTEM

A. Image Processing (Deblurring) Using Raspberry Pi

Image processing is a mechanism to change the image by performing an operation on the image. Image processing is done to get a better picture so that information can be retrieved from the improved image [7]. In this study, the image processing mechanism is performed by using an embedded computer, such as Raspberry Pi. IP camera is used to capture object image and is connected to Raspberry Pi which has an ability to compute image processing.

Raspberry Pi is a small size of computer board which can be used to plug-in a monitor, a keyboard, a mouse, and others [1]. Raspberry Pi has software like scratch which allows the user to program and to design animations, games, or videos [13]. In addition, a programmer can also develop scripts or programs using python language on the Raspbian operating system [12].

In this study, IP camera is going to be connected with Raspberry Pi. Thus image which is caught by IP camera is going to be stored temporarily in Raspberry Pi storage. The blurred image is going to be fixed by image processing mechanism which is deblurring. The deblurring algorithm which is used in this study is the Richardson-Lucy.

B. Deblurring Using Richardson-Lucy

Blurring is a damage which happens to the image which can make the image looks blurry. Generally, it is occurred due to camera movement or due to object movement during shooting [6]. This such blurring is known as motion blur [9] [14].

To fix the damage which is caused by motion blur, the deblurring mechanism is executed. In this study, deblurring is done by using the Richardson-Lucy algorithm.

Here is the formula for the *Richardson-Lucy* algorithm [6]:

$$f^{n+1} = f^n H(\frac{g}{Hf^n}) \quad (1)$$

In formula (1), f^{n+1} is a new estimation of f^n and f is a figure, g is Gaussian blurred image, and H is PSF (Point Spread Function).

Even though Richardson-Lucy is more persistence from noises than other algorithms, there is a limitation that the image must be converted first before doing the image processing mechanism. The image must be converted into a grayscale image. The reason is Richardson-Lucy only can restore two-dimension image layers. It cannot restore more than two dimension image layers like RGB image [17].

C. Point Spread Function (PSF)

Point Spread Function (PSF) is the variable which must be calculated before deblurring. PSF is also known as the diffusion technique of light point [17]. This study uses Gaussian PSF because it can correct blur which is caused by moving object [14]. Here is the formula to get the Gaussian PSF [17]:

$$h_g(m,n) = e^{\frac{-(m^2+n^2)}{2\sigma^2}} \quad (2)$$

$$h(m,n) = \frac{h_g(m,n)}{\sum m \sum n h_g} \quad (3)$$

In formula (2) and formula (3), h is the PSF, (m,n) are pixel values, and σ has a value which equals 1, and e has a value which is 2.71828.

D.Peak Signal to Noise Ratio (PSNR)

Peak Signal to Noise Ratio (PSNR) is used to calculate the similarity between the original image and the restored one. To calculate PSNR requires MSE (Mean Square Error). MSE can be used to calculate the number of errors between the original image and the restored one. Here is the formula for calculating MSE value [14]:

$$\text{MSE} = \frac{1}{M*N}\sum_{x=1}^{M}\sum_{y=1}^{N}(\text{f}(x,y) - \text{f}'(x,y))^2 \quad (4)$$

In formula (4), $M*N$ is the size of the image. Function $\text{f}(x,y)$ dan function $\text{f}'(x,y)$ are pixel values (x,y) of the original image and the restored image. After MSE value collected, then PSNR value can be calculated using the followed formula [14]:

$$\text{PSNR} = 10Log_{10}(\frac{255x255}{\sqrt{MSE}}) \quad (5)$$

The higher PSNR value and the lower MSE value can be determined that the image quality becomes better [14].

E. Google Drive and overGrive

Cloud storage is data storage which data are maintained, managed, and stored. Internet network is required to access the cloud service. In general, the user has to pay for the storage service to get a larger capacity [16]. Even so, the payment mechanism makes its service is cheaper than the cost for maintaining typical server storage [8].

This study uses Google Drive as the cloud storage for saving the image processing result of Surveillance Embedded IP Camera.

The overGrive feature is used to sync image data between Surveillance Embedded IP Camera and the Google Drive [10].

F. *The Components of Hardware and Software*

The hardware required in this study are:

- An embedded computer: Raspberry Pi 3 with local storage (32 GB SD Card).
- A USB Camera: SainSmart Wide Angle Fish-Eye 5 Megapixels (2592 x 1944 pixels).
- A router which is used for connecting between the embedded computer and the cloud system via the internet.

The software used for the embedded computer are:

- Raspbian which acts as Operating System.
- Scikit-Image which is used as computer vision libraries to do image processing which can be used on Mac, Windows, and Ubuntu Linux.
- The overGrive which can be used on Ubuntu Linux and Raspbian. It is used to integrate file and folder on local storage to the cloud storage.

So that the building block of the proposed system can be followed as Fig 1.

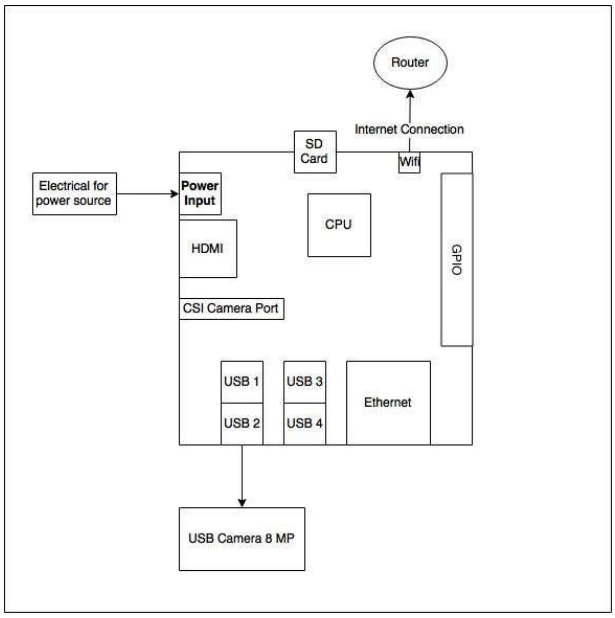


Fig. 1. The building block of the proposed system.

The explanation of the proposed system can be seen on the followed picture:

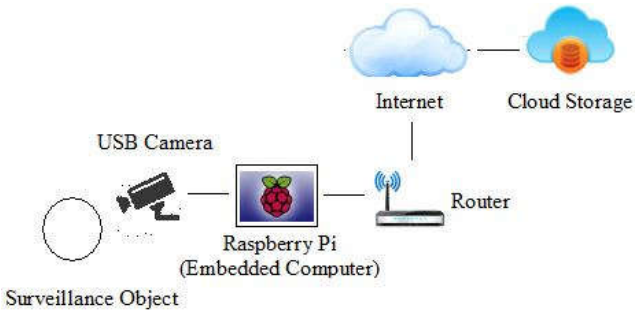


Fig. 2. The proposed system

The object image is captured by a USB camera which is linked into the embedded computer through a USB port. Then the caught picture is saved first in the local memory of the embedded computer and processed for fixing blurred images. The mechanism of the image fixing is done after the captured image is converted into a grayscale image first. Then, the result of the image processing is stored into cloud storage.

V. ANALYSIS AND DISCUSSION

A. *Deblurring*

The deblurring process using the Surveillance Embedded IP Camera is done in 2 steps. The first is to test the captured images of USB camera based on the kernel size. This is done to find the best PSF value. If the best PSF value is already obtained, then the value is going to be used in the next step for deblurring using the Richardson-Lucy method. In the first step, 30 images are tested using different kernel sizes, which are 3 x 3 pixels, 5 x 5 pixels, 7 x 7 pixels, and 9 x 9 pixels. Then the images are going to be deblurred using Richardson-Lucy method.

B. *Finding The PSF Value Based On Kernel Size*

At this stage, 30 images are going to be used to determine the PSF value based on variety kernel sizes. The value then is going to be used for deblurring using Richardson-Lucy method. The several different kernel sizes tested are 3 x 3 pixels, 5 x 5 pixels, 7 x 7 pixels, and 9 x 9 pixels.

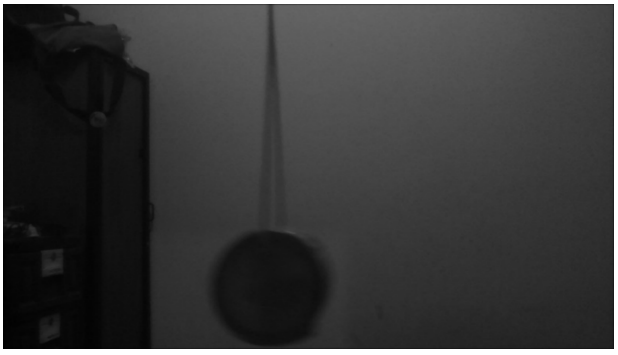


Fig. 3. The examination of PSF using kernel size 3 x 3 pixels.

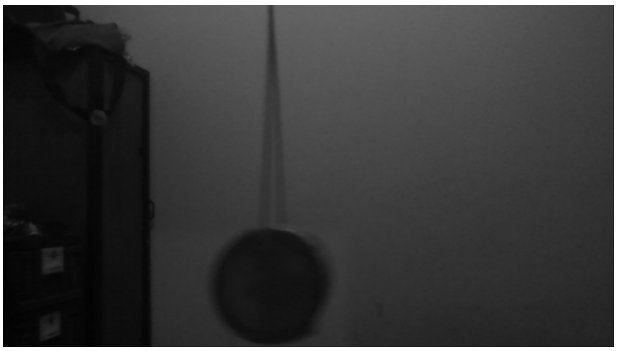


Fig. 4. The PSF examination uses kernel size 5 x 5 pixels.

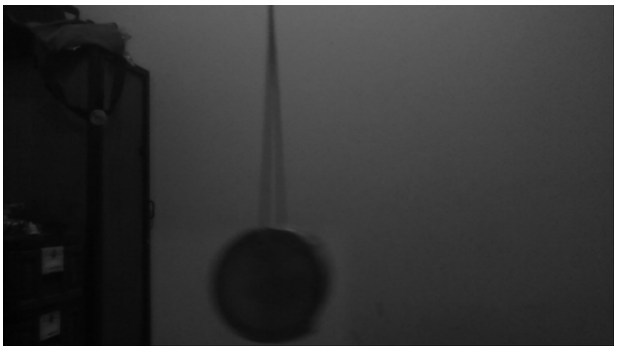


Fig. 5. The PSF examination uses kernel size 7 x 7 pixels.

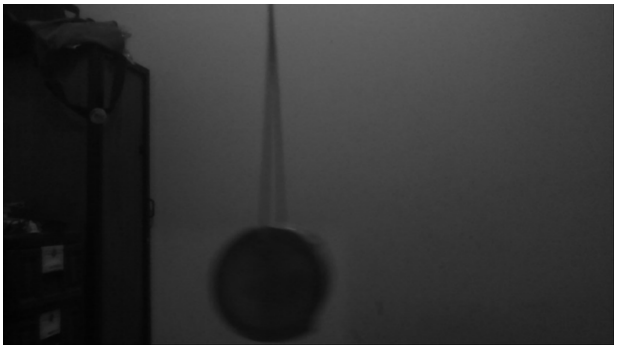


Fig. 6. The PSF examination uses kernel size 9 x 9 pixels.

The followed tables show the PSNR value of 30 tested images using variety kernel sizes (see Table I and Table II):

TABLE I. THE PSNR VALUE OF GRAYSCALE IMAGES.

	The PSNR value of grayscale image (dB)			
	3x3	5x5	7x7	9x9
Total	1336.293709	1348.123132	1354.188292	1354.111945
Average	44.54312362	44.93743772	45.13960975	45.13706483

TABLE II. THE PSNR VALUE OF DEBLURRED IMAGES.

	The PSNR value of *deblurred image* (dB)			
	3x3	5x5	7x7	9x9
Total	1378.681194	1355.243434	1360.998069	1361.062100
Average	45.95603981	45.17478114	45.36660229	45.36873667

Based on the testing, it is found that the kernel size 3 x 3 pixels produces the best PSNR value than other kernel sizes for image deblurring. Then the mechanism of deblurring using Richardson-Lucy method on Surveillance Embedded IP Camera is going to use PSF with kernel size 3 x 3 pixels.

C. Deblurring Using Richardson-Lucy Method

Deblurring is going to be tested using 30 images which are the original images on local storage captured by the USB camera. The Richardson-Lucy method is chosen as the algorithm for deblurring and the PSF value which has been obtained in the previous test, which is kernel size 3 x 3 pixels is adopted. Before deblurring, the original images are going to be converted into grayscale images to be processed using Richardson-Lucy method. Here are some examples used in this step:



Fig. 7. The original image.

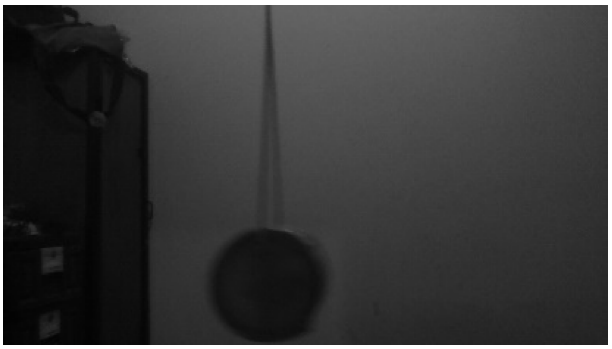


Fig. 8. The grayscale image.

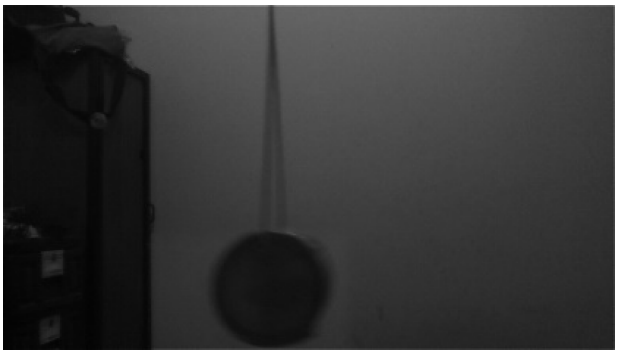


Fig. 9. The deblurred image.

Fig 7 is the original image, Fig 8 is the original image which has been changed to grayscale, and Fig 9 is the image result of deblurring.

To know the quality of the image can be done by comparing the original image and the image which has been through deblurring. After that, both images are going to be searched for its PSNR value, so it can be known the quality of the image. The followed table shows the value of PSNR obtained from 30 tested data:

TABLE III. THE PSNR VALUE OF 30 TESTED DATA.

	The PSNR Value of Grayscale Image (dB)	The PSNR Value of Grayscale Image (dB)
Total	1336.293709	1378.681194
Average	44.54312362	45.95603981

Based on the information of Table III, it shows that the image which is categorized as the original image has a mean value of PSNR as 44.54312362 dB, while the image which is categorized as the deblurred image has a mean value of PSNR as 45.95603981 dB. So that, the deblurred image after deblurring has a better quality than the image before deblurring.

D. Uploading Images to Google Drive Using overGrive

The overGrive is used for sending images from local storage of the embedded computer to cloud storage. It is going to synchronize files and folders on the Raspberry Pi local storage which contains original images and deblurred images, so the images are going to be automatically uploaded to cloud storage, namely Google Drive.

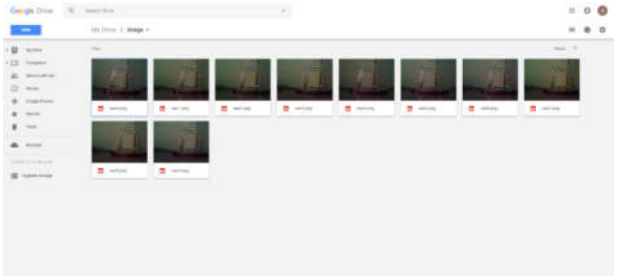


Fig. 10. The images which are successfully uploaded to the cloud storage using overGrive.

The test uses speed around 200 KBps for the 1280 x 720 pixels image size. The purpose of this test is to know that images stored on local storage have been successfully sent to cloud storage by checking the folders in Google Drive. The followed results have been captured can be seen on the followed table:

TABLE IV. THE PSNR VALUE OF 30 TESTED DATA.

No.	The Image Size (KB)	Time Needed for Uploading (Second)
1.	861	4.3
2.	968	4.8
3.	966	4.8
4.	939	4.6
5.	962	4.8
6.	973	4.9
7.	955	4.8
8.	993	5
9.	959	4.8
10.	929	4.6
Total		47.4
Average		4.74

Based on the information of Table IV, it can be seen that the average time required for uploading images to the cloud storage using internet access of 200 KBps is 4.74 seconds.

VI. CONCLUSION

The Surveillance Embedded IP Camera has been developed which has features such as passive surveillance integrated with cloud storage, and image deblurring using Richardson-Lucy method. The PSF testing using 3 x 3 pixels kernel size can generate the best average PSNR value. The use of Richardson-Lucy method with 3 x 3 pixels gives the average PSNR value which is 45.95603981 dB. It is higher than the average PSNR value of the images before deblurring, which is 44.54312362 dB. The difference is 1.41291619 dB. Based on the increase of the average PSNR value shows that the image after deblurring using Richardson-Lucy method has a better image quality than the original grayscale image, although it is not significant. It happens because motion blur is hard to be repaired. The last testing which is the delivery of the images to the cloud storage shows that the average time required when uploading 1280 x 720 pixels images with 200 KBps internet speed is 4.74 seconds.

REFERENCES

[1] A. K. Dennis. Raspberry Pi home Automation with Arduino. Packt Publishing Ltd, 2013.

[2] Amazon.com, "Ctronics Outdoor WIFI Camera Bullet IP Camera HD 1080p Security Wireless Camera 30m Ir 6mm Lens Onvif 2.0 36 Leds Night Vision 16G SD Card Built-in Sony Chipset", amazon.com, 2016. [Online]. Available: https://www.amazon.com/Ctronics-Outdoor-Security-Wireless-Chipset/dp/B01ICNEWLY/ref=sr_1_5?s=photo&ie=UTF8&qid=1478939455&sr=1-5&keywords=Surveillance+Camera+with+Micro+SD+Card+Storage. [Accessed: 9 December 2016].

[3] Amazon.com, "Sinocam 4CH Wifi Wireless 720P NVR HD Security Camera System with 4 720P 1.0MP Weatherproof Superior Night Vision HD IP Security Cameras System", amazon.com, 2016.

[Online]. Available: https://www.amazon.com/dp/B019DX0JTG/ref=asc_df_B019DX0JTG5300542?tag=merbethomgarparenting-20&creative=395261&creativeASIN=B019DX0JTG&linkCode=asn&ascsubtag=250990181. [Accessed: 9 December 2016].

[4] BNP Media, "12 Cost-Efficient Video Surveillance Strategies", securitymagazine.com, 2016. [Online]. Available: https://www.securitymagazine.com/articles/86749-cost-efficient-video-surveillance-strategies. [Accessed: 9 December 2016].

[5] D. a. Rodriguez-Silva, L. Adkinson-Orellana, F. J. Gonz'lez-Castano, I. Armino-Franco, and D. Gonz'lez-Martinez, "Video surveillance based on cloud storage", in Int. Conf. Cloud Comput. (CLOUD), 2012, pp. 991–992.

[6] F. Vankawala, "A survey on different image deblurring techniques", International Journal of Computer Applications. vol. 116, no. 13, pp. 15–18, 2015.

[7] G. Anbarjafari. "Digital Image Processing", University of Tartu. 2014. [Online]. Available: https://sisu.ut.ee/imageprocessing/book/1. [Accessed 20 December 2017].

[8] Google, "Google Drive storage plans & pricing - Drive Help", support.google.com, 2016. [Online]. Available: https://support.google.com/drive/answer/2375123?hl=en. [Accessed: 9 December 2016].

[9] J. Wu, C. Chang, and C. Chen, "An adaptive Richardson-Lucy algorithm for single image deblurring using local extrema filtering", vol. 16, no. 3, pp. 269–276, 2013.

[10] Ohso, "OverGrive Is a Google Drive Desktop Client for Linux", omgubuntu.co.uk, 2016. [Online]. Available: http://www.omgubuntu.co.uk/2016/10/overgrive-google-drive-desktop-client. [Accessed: 27 October 2016].

[11] P. Natarajan, "Iterative Richardson-Lucy based improved fast image deblurring and denoising approach", International Journal for Advance Research in Engineering and Technology (ICIRET). pp. 2320-6802, 2014.

[12] Raspberry Pi Foundation, "Phyton", raspberrypi.org, 2017. [Online]. Available: https://www.raspberrypi.org/documentation/usage/python/. [Accessed: 01 Agustus 2017].

[13] S. McManus, and M. Cook. Raspberry Pi for Dummies. John Wiley & Sons, 2017.

[14] S. Sharma, S. Sharma, and R. Mehra, "Image restoration using modified Lucy Richardson algorithm in the presence of Gaussian and motion blur", Advance in Electronic and Electric Engineering, vol. 3, no. 8, pp. 1063–1070, 2013.

[15] S. Singh and P. J. P. Chavan, "IP camera video surveillance using Raspberry Pi", International Journal of Advance Research in Computer and Communication Engineering, vol. 4, no. 2, pp. 326–328, 2015.

[16] TechTarget, "Definition Cloud Storage", techtarget.com, 2016. [Online]. Available: http://searchcloudstorage.techtarget.com/definition/cloud-storage. [Accessed: May 2016].

[17] Z. Al-Ameen, G. Sulong, M.G.M. Johar, N. Verma, R. Kumar, M. Dachyar, M. Alkhawlani, A. Mohsen, H. Singh, and S. Singh, "A Comprehensive Study on Fast Image Deblurring Techniques", International Journal of Advanced Science and Technology, vol. 44, 2012.

Supplier Selection Using Combined Method of K-Means and Intuitionistic Fuzzy Topsis

Melany Mustika Dewi
Master of Informatics Engineering
AMIKOM University
Yogyakarta, Indonesia
melany.m@students.amikom.ac.id

Kusrini
Master of Informatics Engineering
AMIKOM University
Yogyakarta, Indonesia
kusrini@amikom.ac.id

Hanif Al Fatta
Master of Informatics Engineering
AMIKOM University
Yogyakarta, Indonesia
hanif.a@amikom.ac.id

Abstarct -- Supplier selection as an important subject for survival of a business enterpris life. The selection of appropriate suppliers can be assisted by decision support methods. This research with title "Supplier Selection Using Combined Method of K-Means and Intuitionistic Fuzzy Topsis" purposed for help giving right supplier recomendation using decision making application using K-Means and Intuitionistic Fuzzy Topsis method at Citra Mart University Amikom Yogyakarta. In this research supplier decision making criteria are using price, delivery order, delivery speed, stock availibility, quality of service and purchase method. Research data is using 6 goods with 13 suppliers. This research is looking for the best supplier from many supplier of one goods with using combined Intuitionistic Fuzzy Topsis and K-Means for help clustering data with number value. In Fuzzy Topsis method is using eulidean distance algorithm for fixing Intuitionistic Fuzzy Set (IFS) number. Research is begin with diagnosing problem, system analysis, implementation to be application and evaluating result of method. Decision making evaluating using Mean Absolute Error (MAE). The result of this research has accuracy level of 0,64 from clustering, so this result can be concluded that succesfully to make a right price clustering and succesfully giving a right recomendation of supplier selection with using that method.

Keywords—Supplier Selection; K-Means; Clustering; Intuitionistic Fuzzy Topsis; Decision Making;

I. INTRODUCTION

Sulistyani (2017) and Heizer (2017) said Supply Chain Management (SCM) is a global relationship between suppliers, manufacturers, distribution centers, wholesalers, retailers and end-consumers in the process of supplying goods and services so that products can be distributed with the right amount and location, minimal cost and high customer satisfaction [1] [2]. In supply chain management, suppliers are a very important part and affect the viability of a business organization [1].

The selection of suppliers that match the demands of business organizations is important. Suppliers are third parties who guarantee the availability of goods required by business organizations. If the delivery of goods is late there will be a void of goods and will decrease customer demand, and if the quality of goods does not meet the standards it will lose the customer's trust as well.

Suppliers are selected based on specified criteria in order to get the right supplier. Determining the right suppliers can solve problems in supplier selection. Criteria suppliers to be used also in accordance with the object to be researched because each object has its own characteristics.

The selection of appropriate suppliers can be done with the help of decision support system to produce one of the best suppliers and in accordance with the needs of the business organization. When using multiple suppliers for a single item will be difficult in warehouse inventory and administration. Decision support system is a system that can provide advice to decision makers in decision-making but not to replace decision makers.

In this research using K-Means method for clustering and combination of Intuitionistic Fuzzy Set and Topsis for decision making method. The Topsis method is easily for measuring the performance of alternatives with simple computing and can rank preference values based on proximity positive and negative ideal solutions. Other researchers used Topsis method to determine the spread of the journal [3], to determine selection of drugs on RSIA Tiara [4], Topsis used for shopping engines as a buyer's decision support in choosing goods [5], Topsis is also combined with other methods to select raw material suppliers for small and medium enterprises [6] . This method requires subjective input of the weight of the decision maker. To facilitate the input weight, then used intuitionistic method fuzzy set by using linguistic scale as

input. The intuitionistic fuzzy set (IFS) method is used to overcome doubts and uncertainties in the human decision-making process. The Topsis method when combined with intuitionistic fuzzy sets has a high success rate in overcoming Multi Criteria Decision Making (MCDM) problems due to the uncertainty in the opinions given when making decisions and generating ratings consistent with alternatives and criteria.

The combination of IFS and Topsis methods in its development is combined with other methods in different areas of research such as the engine troubleshooting of shipboard damage using IF- Topsis [7], the field of quality of service such as determining the best service quality of all airports in North African as well using IF- Topsis [8], the business field for supplier selection issues using IF-Axoimatic design (IFAD) [9]. IFS is also incorporated into other methods such as Regular Increasing Monotone (RIM) and Gaussian methodare utilized in the search for weightings in journals [10], and using for suppliers selection in the development of petroleum industry facilities [11].

IFS is used in assigning weighted values to criteria as well as to alternative suppliers. Giving weight to this method uses linguistic criteria such as high, low, average. The use of these linguistic criteria facilitates decision makers in weighing the criteria data. The use of linguistic criteria is also used by other researchers in the selection of supplier using the Fuzzy Topsis method and probabilistic criteria [12]. In one criterion there are two weights at the same time representing the doubt of a human being and worth less than one. IFS has a value that matches human estimates as good or very good so it will be difficult when the value to be processed is numerical or linear data type like price. The solution of this problem is to grouping numerical data type into linguistic criteria by clustering process.

Clustering is one of the techniques in data mining that aims to data classify that has the same characteristics into one group. Some of the clustering methods are K-Means, Single Linkage, Fuzzy C Means and K-Medoids. In this study using K-Means method that can classify data based on maximum resemblance and maximal inclination in one cluster or group based on predetermined cluster number. With this grouping it is hoped that there are no clusters with differences in distance too far or too close between linguistic criteria with each other. The K-Means method will be used to solve the numerical data problem to be converted to linguistic criteria. Other researchers also used the cluster method of grouping sales by using the same method of K-Means [13]. Based on the background that has been described, the writer tries to apply the K-Means method as a cluster method for the price criteria into linguistic and intuitionistic fuzzy Topsis method using for ranking process of decision-making.

Researcher using Mean Absolute Error (MAE) for calculate accuration from system ranking with user ranking. MAE can gives result with equal accuration level for all error, so can gives more accuration result.

II. METHODOLOGY

This study presents a combination of K-Means and intuitionistic fuzzy Topsis methods to find the right supplier selection. First step is diagnosing problem with searching literatur study and looking for research object to collect data and knowing more about the problem. This research uses supplier selection in Citra Mart as research object. After that, choosing method to solving problem and then make a research methodology. The research methodology of supplier selection in Citra Mart will be described in this chapter.

A. Research Methodology

In Fig.1 explains the research methodology in the form of flowcharts.

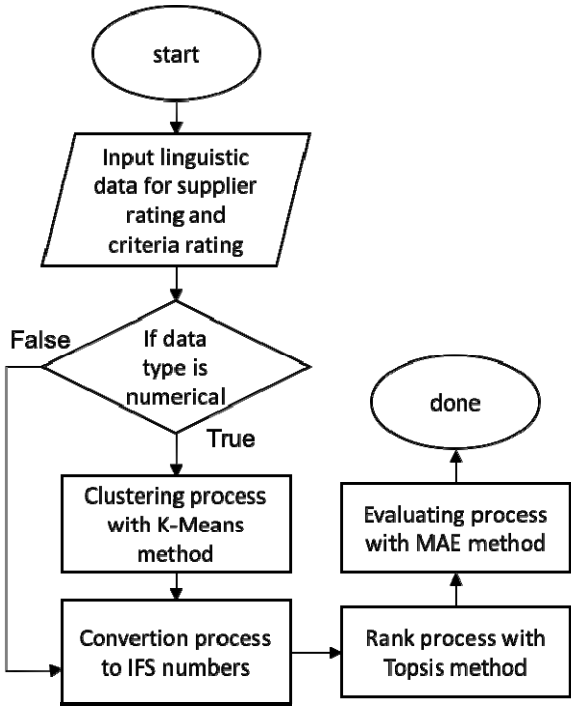


Fig. 1. Research Methodology

Step 1: Analyze the data requirements for supplier selection likes criteria, goods data and supplier data.

Step 2: Determine the linguistic scale for data rating with IFS numbers.

Step 3: Assess the weighting criteria and assessment of alternative criteria for suppliers from a single item using a predetermined linguistic scale.

Step 4: If the data assessment is a number then grouped first into one group of linguistic scale using K-Means cluster method.

Step 5: Convert the scoring result from the lingusitic scale to the IFS number.

Step 6: Determine the ranking of supplier alternatives using the fuzzy Topsis method.

III. DISCUSSION AND RESULT

A. Data Requirements

In this chapter will present the discussion of the methodology already described. Here is a sample calculation for the selection suppliers of Citra Mart. The used criteria are Price (C1), Quality (C2), Delivery (C3), Purchase Method (C4). For criteria C2 is divided into two namely the quality of service and the quality of goods avaibility. Delivery criteria also divided into two namely the delivery order and delivery speed. For more clearly can be seen the hierarchy diagram criteria in Fig.2. There are 4 main criteria with two nested criteria on the main criteria considered in this study.

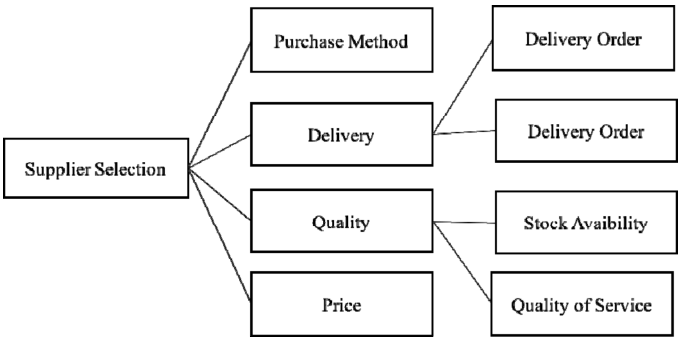


Fig. 2. The criteria hierarchy diagram of supplier selection at Citra Mart

B. Determination of linguistic scale

The linguistic scale for the assessment of criteria and assessment of supplier alternatives can be seen in Table I and Table II. The linguistic scale is the scale that helps determine the value of the criterion as well as the suppliers of alternative suppliers in the form of linguistics or reasoning. The value of the linguistic scale is intuitionistic intuitionistic set (IFS). This IFS number weighting will be used in the next ranking process.

TABLE I. LINGUISTIC SCALE FOR WEIGHTING CRITERIA ASSESSMENT

Scale		IFS Numbers
VI	Very Important	(0.9,0.1)
I	Important	(0.6,0.2)
N	Neutral	(0.4,0.3)

TABLE II. THE LINGUISTIC SCALE FOR THE ASSESSMENT OF ALTERNATIVE CRITERIA

Scale		IFS Numbers
VG	Very Good	(0.9,0.1)
G	Good	(0.6,0.2)
F	Fair	(0.4,0.3)

C. Providing the value of alternative weights and assessment of suppliers

Assessment of supplier criteria and alternatives using linguistic scales is tabulated in Table III. Assessment uses initialization of the lingusitic scale.

TABLE III. WEIGHTING AND ALTERNATIVE ASSESSMENT OF SUPPLIERS

	C1	C2	C3	C4
RATING/ SUPPLIER	VI	I,I	I,I	I
A1	13.000	G	F	G
A2	12.000	G	F	G
A3	13.000	G	F	G
A4	12.000	G	VG,B	VG
A5	11.500	VG	VG,B	VG
A6	11.500	VG	VG,B	VG
A7	12.500	VG,F	F	G
A8	11.500	VG,F	VG,G	VG
A9	13.000	VG,F	F	G
A10	13.000	G,F	F	G

D. Data clusters with K-Means

In the valuation data presented in Table III there is a price with numerical data type then do the cluster process using K-Means to group the price into one linguistic scale. Here is the step of the cluster process.

STEP 1: Determine the number of clusters to be created. Linguistic scale for the price there are two that is Very Good and Good. Only two scales are given because the suppliers are picked up only at the price that best fits the object. The number of clusters to be made according to the lingusitic scale for the price is two. Cluster 1 for Very Good scale (VG) and Cluster 2 for good scale (G) to further be abbreviated as K1 for cluster 1 and K2 for cluster 2.

STEP 2: Initialize k cluster center. The center cluster or called the centroid is the average value of all data on a particular cluster. Centroid is given a unique and different initial value of each cluster can by using the maximum or minimum value of all data. In equations 1 and 2 the formula for determining the centroid with two clusters of clusters.

$$cK1 = Min(H_1,H_2,H_3,H_4,H_5,....H_n) \quad (1)$$

$$cK2 = Max(H_1,H_2,H_3,H_4,H_5,....H_n) \quad (2)$$

With cK1 and cK2 is the centroid value of each cluster and H_1-H_n is the set of price data to be clustered.

cK1=Min(13000,12000,13000,12000,11500,11500,12500, 11500,13000,13000) = 11500

cK2=Max(13000,12000,13000,12000,11500,11500,12500, 11500,13000,13000) = 13000

STEP 3: Define all data to the nearest cluster. To calculate the distance of all data to each cluster center can use Euclidean distance theory in equation 3. Subsequently the result of calculation of supplier distance is presented in Table IV.

$$D(i,j) = \sqrt{(H1i\text{-}H1j)^2 + \cdots + (Hki\text{-}Hkj)^2} \quad (3)$$

Where :

D (i,j) = Distance of data to i to centroid j

Hki = Data i on attribute data to k

Hkj = Centroid to j on attribute i

TABLE IV. CALCULATION OF DISTANCE SUPPLIERS WITH CENTROID

	Distance to K1	Distance to K2	Result
A1	1500	0	K2
A2	500	1000	K1
A3	1500	0	K2
A4	500	1000	K1
A5	0	1500	K1
A6	0	1500	K1
A7	1000	500	K2
A8	0	1500	K1
A9	1500	0	K2
A10	1500	0	K2

STEP 4 : Determine the centroid based on the cluster results from Step 3. The centroid determination is done by summing all the data on the same cluster divided by the amount of data in the cluster using equations 4 and 5. Table V presents the calculation of the new centroid based on the results of the distance calculation Table IV .

$$cK1n = Sum(H_1 \ldots H_n \ K1)/ \bullet n \quad (4)$$

$$cK2n = Sum(H_1 \ldots H_n \ K2)/ \bullet n \quad (5)$$

cK1n = Sum(12000,12000,11500,11500,11500)/5

= 58500/5 = 11700

*cK2*I*n* = Sum(13000,13000,12500,13000,13000)/5

= 64500/5 = 12900

STEP 5 : Define again all data into the cluster by using the new cluster center. If the cluster results do not change then the clustering process is complete. Alternatively, return to step number 3 until the center of the cluster does not change anymore. The result of grouping price data into linguistic scale is presented in Table V. Scale for criteria K1 is Very Good and K2 is Good.

TABLE V. THE RESULTS ARE GROUPED INTO LINGUISTIC SCALES

	Distance to K1	Distance to K2	Result	LS
A1	1300	100	K2	G
A2	300	900	K1	VG
A3	1300	100	K2	G
A4	300	900	K1	VG
A5	200	1400	K1	VG
A6	200	1400	K1	VG
A7	800	400	K2	G
A8	200	1400	K1	VG
A9	1300	100	K2	G
A10	1300	100	K2	G

E. Convertion of IFS Numbers

Assessment of the criteria weight and supplier alternatives in the form of linguistic scale are converted into IFS numbers according to Table I and Table II. The following is the result of the conversion of criteria criteria and alternative criteria in Table VII.

TABLE VI. CONVERSION RATING TO IFS NUMBER

	Criteria			
	C1 (0.9,0.1)	C2 (0.6,0.2), (0.6,0.2)	C3 (0.6,0.2), (0.6,0.2)	C4 (0.6,0.2)
A1	(0.6,0.2)	(0.6,0.2)	(0.4,0.3)	(0.6,0.2)
A2	(0.9,0.1)	(0.6,0.2)	(0.4,0.3)	(0.6,0.2)
A3	(0.6,0.2)	(0.6,0.2)	(0.4,0.3)	(0.6,0.2)
A4	(0.9,0.1)	(0.6,0.2)	(0.9,0.1), (0.6,0.2)	(0.9,0.1)
A5	(0.9,0.1)	(0.9,0.1)	(0.9,0.1), (0.6,0.2)	(0.9,0.1)
A6	(0.9,0.1)	(0.9,0.1)	(0.9,0.1), (0.6,0.2)	(0.9,0.1)
A7	(0.6,0.2)	(0.9,0.1), (0.4,0.3)	(0.4,0.3)	(0.6,0.2)
A8	(0.9,0.1)	(0.6,0.2), (0.4,0.3)	(0.9,0.1), (0.6,0.2)	(0.9,0.1)
A9	(0.6,0.2)	(0.9,0.1), (0.4,0.3)	(0.4,0.3)	(0.6,0.2)
A10	(0.6,0.2)	(0.6,0.2), (0.4,0.3)	(0.4,0.3)	(0.6,0.2)

F. Rank Process with Fuzzy Topsis Method

From the data that has been converted to the next IFS number is to determine the selection of suppliers with the help of fuzzy Topsis method with the following steps:

STEP 1: Create a fuzzy matrix decision. For ease of calculation, the IFS number is made into a single value using the euclidean distance formula in equation 6.

$$N(i,j) = (i\text{-}max(i)) * 2 + (j\text{-}min(j)) * 2 + ((1 - max(i) + min(j)) - ((1 - (max(i) - min(j)) * 2 \quad (6)$$

N(0.9,0.1) = (0,9-0,9)*2 + (0,1-0,1)*2 + ((1-(0,9+0,1)) - (1-(0,9-0,1)))*2 = 0,04

N(0.6,0.2) = (0,6-0,9)*2 + (0,2-0,1)*2 + ((1-(0,6+0,2)) - (1-(0,9-0,1)))*2 = 0,1

N(0.4,0.3) = (0,4-0,9)*2 + (0,3-0,1)*2 + ((1-(0,4+0,3)) - (1-(0,9-0,1)))*2 = 0,3

With a value inversely proportional to the actual value of the formula is added a new formula to find the value of the opposite by using the formula in equation 7.

$$d = 1 - n \quad (7)$$

d(0.9,0.1) = 1 − 0.04 = 0.96

d(0.6,0.2) = 1 − 0.1 = 0.9

d(0.4,0.3) = 1 − 0.3 = 0.7

For d is the divergen value or the inverse value and N is the single value of the result with the euclidean distance. For an assessment of two values, the result of a single value is divided by two as in eq.8.

$$Cx = (n_a + n_b) / 2 \quad (8)$$

$$C = \begin{bmatrix} 0,9 & 0,9 & 0,7 & 0,7 \\ 0,96 & 0,9 & 0,7 & 0,7 \\ 0,9 & 0,9 & 0,7 & 0,7 \\ 0,96 & 0,9 & 0,83 & 0,96 \\ 0,96 & 0,96 & 0,83 & 0,96 \\ 0,96 & 0,96 & 0,83 & 0,96 \\ 0,9 & 0,45 & 0,7 & 0,7 \\ 0,96 & 0,48 & 0,83 & 0,96 \\ 0,9 & 0,48 & 0,7 & 0,7 \\ 0,9 & 0,45 & 0,7 & 0,7 \end{bmatrix}$$

STEP 2 : Create a fuzzy normalization of decision matrix. Calculate the normalization of fuzzy decision matrix by using eq.9.

$$c_{ij} = \frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}^2} \quad (9)$$

Where :

C_{ij} = result of C fuzzy normalization of decision matrix

i = 1,2,3,4.....m ; j = 1,2,3,4…..,n

$$C = \begin{bmatrix} 0,36308 & 0,30786 & 0,29331 & 0,27193 \\ 0,38729 & 0,30786 & 0,29331 & 0,27193 \\ 0,46308 & 0,30786 & 0,29331 & 0,27193 \\ 0,38739 & 0,30786 & 0,34778 & 0,37294 \\ 0,38739 & 0,32828 & 0,34778 & 0,37294 \\ 0,38739 & 0,32828 & 0,34778 & 0,37294 \\ 0,38739 & 0,15393 & 0,29331 & 0,27193 \\ 0,36308 & 0,16419 & 0,34778 & 0,37294 \\ 0,38739 & 0,16419 & 0,29331 & 0,27193 \\ 0,36308 & 0,15393 & 0,29331 & 0,27193 \end{bmatrix}$$

STEP 3 : Create a weighted matrix decision fuzzy using the formula in equation 10.

$$C_{ij} = x_i r_{ij} \quad (10)$$

$$C = \begin{bmatrix} 0,34856 & 0,27707 & 0,26398 & 0,24474 \\ 0,7180 & 0,27707 & 0,26398 & 0,24474 \\ 0,34856 & 0,27707 & 0,26398 & 0,24474 \\ 0,37180 & 0,27707 & 0,31301 & 0,33564 \\ 0,37180 & 0,29554 & 0,31301 & 0,33564 \\ 0,37180 & 0,29554 & 0,31301 & 0,33564 \\ 0,34856 & 0,13854 & 0,26398 & 0,24474 \\ 0,37180 & 0,14777 & 0,31301 & 0,33564 \\ 0,34856 & 0,14777 & 0,26398 & 0,24474 \\ 0,34856 & 0,13854 & 0,26398 & 0,24474 \end{bmatrix}$$

STEP 4 : Calculate the ideal ideal fuzzy solution and fuzzy ideal negative solution. The ideal ideal solution and the ideal negative solution can be determined with a normalized weighted rating as in equations 11 and 12. For more details see Table VIII.

$$A^+ = (y_1^+ , y_2^+ , \ldots . y_n^+), \quad (11)$$

$$A^- = (y_1^- , y_2^- , \ldots . y_n^-), \quad (12)$$

with;

y+ = max y*ij* (benefit) ; min y*ij* (cost)

y-= min y*ij* (benefit);max y*ij* (cost)

TABLE VII. RESULT OF FSIP DAN FSIN

	C1	C2	C3	C4
A+	0,34856	0,29554	0,31301	0,33564
A-	0,37180	0,13854	0,26398	0,24474

STEP 5 : Calculate the proximity coefficient. The distance between alternative Ai with positive ideal solution and negative ideal solution is formulated uses eq.13 and eq.14 :

$$D_i^+ = \sqrt{\sum_{j=1}^{n} (y_i^+ - yij)^2} \quad where\ i = 1,2,.....m \quad (13)$$

$$D_i^- = \sqrt{\sum_{j=1}^{n} (yij - y_i^-)^2} \quad where\ i = 1,2,.....m \quad (14)$$

The preference value for each alternative (Vi) uses eq.15.

$$Vi = \frac{D_i^-}{D_i^- - D_i^+} \quad where\ i = 1,2,3,.....m \quad (15)$$

STEP 6 : Determine the rankings based on the results of the proximity coefficient. In Table VIII the results of the supplier selection ranking in Citra Mart and Fig.3 show the coefficient closeness result based on the ranking of the supplier alternatives.

TABLE VIII. RANK RESULT

	D+	D-	CC	Rank
A5	0,03873	0,20939	0,843903269	1
A6	0,03873	0,20939	0,843903269	2
A4	0,04383	0,19262	0,814631178	3
A2	0,12284	0,15471	0,567540808	4
A1	0,11748	0,15418	0,567540808	5
A3	0,11748	0,15418	0,557399775	6
A10	0,20932	0,00872	0,40798015	7
A8	0,16869	0,11625	0,062820521	8
A9	0,20084	0,01346	0,039972324	9
A7	0,20932	0,00872	0,039972324	10



Fig. 3. Rank graph of supplier selection

G. Evaluation

From the ranking results in Table VIII an evaluation is conducted by comparing the results of supplier selection with a supplier selection ranking of the object. The result of comparison of the ranks between the results of the method and the object is presented in Table IX.

TABLE IX. COMPARATION OF RANK RESULT

Rank	1	2	3	4	5	6	7	8	9	10
KIFT	A5	A6	A4	A1	A3	A2	A8	A9	A7	A10
Object	A5	A6	A4	A2	A3	A1	A7	A8	A9	A10

Evaluation of this ranking using the formula MAE as in equation (16). The results has accuracy level of 60% so can concluced that using of this combined method is successfully.

$$MAE = \frac{count\ of\ same\ rank}{count\ data} * 100\% \qquad (16)$$

$$MAE = \frac{6}{10} * 100\% = 60\%$$

IV. CONCLUSION

The result of research for supplier selection in Citra Mart using combination of K-Means method and Intuitionistic Fuzzy Topsis can recommend alternative supplier that match with object condition with accuracy level of 60%. For the next researcher can use the method for research in the field and other objects or can develop methods already used with the same object.

REFERENCES

[1] E. Sulistyani, M.I.H. Amir,K.R. Yusuf, N. Nasrullah and D. Injarwanto, "Implementasi Metode Analytical Hierarchy Process (AHP) Sebagai Solusi Alternatif Dalam Pemilihan Supplier Bahan Baku Apel Di PT. Mannasatria Kusumajaya", Technology Science and Engineering Journal, E-ISSN: 2549-1601, Volume 1, No 2, Juni 2017

[2] J, Heizer, B. Render, and C. Munson, Operational Management Sustainability and Supply Chain Management, Pearson, 2017

[3] K. Kusrini, "Paper Distribution Decision for Multi-Journal System", Asian Journal of Information Technology 12(15): 2549-2554, Medwell Journals, ISSN : 1682-3915, 2016

[4] AA. Larasati, AH. Setyaningrum and LK. Wardhani, "Development Decision Support System of Choosing Medicine using TOPSIS Method Case Study : RSIA Tiara" , 6th International Conference on Information and Communication Technology for The Muslim World, 2016

[5] R. Sadida, FH. Purwanto, MM. Dewi and K. Kusrini, "PERANCANGAN SHOPPING SEARCH ENGINE SEBAGAI SISTEM PENUNJANG KEPUTUSAN PEMBELI BERDASARKAN REVIEW DAN HARGA", Seminar Nasional Teknologi Informasi dan Multimedia, *ISSN : 2302-3805, 2018*

[6] S.Chaising and P. Temdee, "Application of a Hybrid Multi-Criteria Decision Making Approach for Selecting of Raw Material Supplier for Small and Medium Enterprises", IEEE, 978-1-5090-5210-3/17, 2017

[7] D.O. Aikhuele, S. Shahryar, H.A. Richard and F.M. Turan, "Application of Intuitionistic Fuzzy Topsis Model for Troubleshooting an Offshore Patrol Boat Engine", Polish Maritime Research 2 (94), Vol. 24, pp. 68-76 10.1515/pomr-2017-0051, 2017

[8] T. Houssine and R. Abdelwaheb, "A TOPSIS method based on intuitionistic fuzzy values: a case study of North African airports", Management Science Letters 7 351-358, 2017

[9] B. Gulcin and G.Fethullah, "Application of a new combined intuitionistic fuzzy MCDM approach based on axiomatic design methodology for the supplier selection problem", Applied Soft Computing Journal, http://dx.doi.org/10.1016/j.asoc.2016.08.051,2016

[10] Thiagarasu and R.Dharmarajan, "An Intuitionistic Fuzzy Topsis DSS Model with Weight Determining Methods ",International Journal Of Engineering And Computer Science, ISSN: 2319-7242 Volume 6 Issue 2 Feb. 2017, Page No. 20354-20361•, Index Copernicus Value (2015): 58.10, DOI: 10.18535/ijecs/v6i2.34

[11] D.A. Wood, "Supplier selection for development of petroleum industry facilities, applying multi-criteria decision making techniques including fuzzy and intuitionistic fuzzy TOPSIS with flexible entropy weighting", Journal of Natural Gas Science and Engineering 28, 2016

[12] S. Bandyopadhyay, "Ranking of Suppliers with MCDA Technique and Probabilistic Criteria", International Conference on Data Science and Engineering (ICDSE), 2016

[13] K.Kusrini, "Grouping of Retail Items by Using K-Means Clustering", Procedia Computer Science 72 (2015) 495-502), 20

Heart Rate Monitoring System in Virtual Environment: The Case of Daily Exercise with Virtual Reality

Abas Setiawan*, Faik Agiwahyuanto†, Pramudi Arsiwi‡
*Informatics Engineering Departement, Computer Science Faculty
Email: abas.setiawan@dsn.dinus.ac.id
†Medical Record and Health Information Departement, Health Science Faculty
Email: faik.agiwahyuanto@dsn.dinus.ac.id
‡Industrial Engineering Departement, Engineering Faculty
Email: pramudi.arsiwi@dsn.dinus.ac.id
Dian Nuswantoro University, Semarang, Indonesia

***Abstract*—The emergence of impressive virtual reality (VR) devices make the virtual reality applications are very popular. There are several VR applications which is involve the user's physical activity while making any interaction with virtual entities. Sometimes the user is not aware that physical activity can drain a lot of energy. The effect of the user who lacks energy is user feel exhausted. One indicator of exhausted is heart rate changes. Current wearable devices have a lot of attached sensor. We proposed the model to monitor users heart rate while making any physical activities in the virtual environment by using the wearable device. Machine to Machine (M2M) communication through an internet network (wireless) can be used to integrate the VR system and wearable device system. In this case, user simulate daily exercise through VR application to make any physical activities. Users felt that this heart rate monitoring system was satisfied, effective, and efficient measured by an adapted System Usability Scale (SUS) questionnaire.**

***Index Terms*—virtual environment, heart rate, wearable device, physical activity, SUS**

I. Introduction

Virtual reality (VR) is a technology that allows users to interact with the virtual environment [1]. Currently, VR technology is very popular because of impressive VR devices such as Oculus Rift and HTC Vive. Oculus Rift and HTC Vive have a Head Mounted Device (HMD) as VR display device. Smartphones which have inertial sensors (accelerometer and gyroscope) can also try this technology with a VR headset or Cardboard equipped.

VR system consists of user input, application, rendering, and output to the user [1]. Input obtained from user response in a virtual environment. The response is button pressed in HMD or controller and where the users gaze located. The application contains a non-rendering aspect of a virtual environment that organize geometric updating, interaction, and physical simulation. Rendering is the transformation of virtual objects from a computer-friendly format into a user-friendly format such as visual rendering, auditory rendering, and haptic rendering. The output is a physical representation that users receive directly through HMD [1]. VR applications have been developed in various areas such as digital games [2], health [3], cultural heritage preservation [4], education [5], and therapy rehabilitation [6].

There are several VR applications which is involve the user's physical activity while making any interaction with virtual entities. Sometimes user felt the time become faster while making physical activities in a virtual environment. The user is also not aware that physical activity can drain a lot of energy. Hence lack of energy cause user feels exhausted. One indicator of exhausted is heart rate changes.

A person who regularly performs physical activity has a quick blood flow through the blood vessels. Thus, the rapid blood flow is affected by the hormone adrenaline and stimulates the nerves to move the heart to pump blood quickly. Hence, blood pressure will also increase and train blood vessels to be elastic. Good elasticity of blood vessels will make a person avoid heart disease [7], [8].

Examination of the person's heart rate conducted in several methods, namely manually and medical equipment assistance. The manual heart rate examination uses the palpation method in the brachial vein, jugular vein, and tricuspid vein for one minute. One minute examination conducted by palpation for 10 seconds multiplied by 6, or 15 minutes multiplied by 4. Furthermore, the result calculated in 60 seconds [9], [10].

The emergence of input devices which have a heart rate sensor allows users to be able to see the heart rate in real-time. One such device is a smartwatch. Smartwatch is a kind of the wearable devices. A wearable device is a machine who attached to the user's body. Even though there is a wearable device, a user who is in a virtual environment cannot see their heart rate directly. The only way to see the heart rate on a wearable device, the user must release the HMD first. A wearable device has a variety of sophisticated features. One of the sophisticated features is M2M communication which it enabled with wireless networks. M2M communications utilize wired or wireless network communications that embedded on a device to communicate with other devices [11].

By utilizing M2M communications, it is possible to monitor the user's heart rate through a smartwatch device while performing activities in the virtual environment. In this research, we will use case studies on VR apps for physical exercise that are used to force the user doing physical activity in virtual environments.

II. Literature Review

A. Heart rate sensor on wearable device

Heart rate is one of the indicators of measuring vital signs that occur in humans. Vital signs will always change at any time, not only in each day but also in every second will always change followed by physical activity. Vital signs on the human body are interrelated. When there is one organ disturbed, then other organs will also be disrupted. One of the vital signs is the rise and fall of the heart rhythm.

An important thing to reduce the incidence of cardiovascular disease (illness or death) is doing exercise. Heart rate is mediated directly by autonomic nervous activity. Where the nerve work is directly affected by the associated hormone gland are adrenal gland of the cortex that produces the adrenal hormone or epinephrine [12].

A person who regularly performs physical activity will have blood flowing rapidly through the blood vessels. So it will affect the hormone adrenaline and stimulate the nerves to move the heart to pump blood quickly. Blood pressure will also increase and train blood vessels to become more elastic. Thus, a person will be able to avoid heart disease if it has elastic blood vessels [7], [8].

The development of wearable sensors to monitor health is a particular concern recently. Several sensors in health monitoring systems on wearable devices can measure the parameters such as heart rate, blood pressure, body and skin temperature, oxygen saturation, respiration rate, electrocardiogram, etc. The purpose of embedding these sensors is to make the user can perform early detection of health conditions experienced [13]. Today, many companies are starting to make wearable devices such as, smart watches, band, headsets, and other devices that have health sensors embedded in it.

B. M2M Communication

Machine to Machine (M2M) Communication is an emerging term in the industry and is part of the Internet of Things (IoT). M2M allows communication between two or more devices in a wired or wireless network to collect information or exchange data [11]. The wired or wireless network usually embedded in hardware that acts as a data acquisition machine or data exchange.

M2M communication technology is a field that will continue to grow in the industrial sector nowadays. Devices such as Apple Watch not only able to tell time and to provide restaurant recommendations. They are also equipped with sensors with such capabilities to monitor activity, capture data, and report information back to other devices or apps without much human intervention [14]. Other devices such as Samsung Gear, MisFit Shine, Jawbone UP, and Xiaomi Mi Band can also be used to track activity and monitor heart rate accurately and precisely [15].

C. Virtual Reality for physical activities

VR applications for physical activity simulations allow the user to interact with 3D objects in a virtual environment [16], [17]. It makes users feel surprised because they can engage in a full and communicate with existing virtual environment. One of the physical activity simulations is a sport. Jinzhao W proposed VR application for competitive sports simulations [18]. Athletes can compete against virtual enemies. The virtual enemies are stronger than the user. The virtual enemies data obtained from existing athletes. Thus, athletes who practice with this VR will have the advantage to learn new tactics, gain psychic strength, and physical training.

Similarly, Vignais N et al created a VR system to analyze the situation during exercise [19]. The situational analysis conducted by the handball goalkeeper while catching the ball. There are two tasks that goalkeeper must do: the goalkeeper shows the ball stops at the video clip and the goalkeeper try to intercept ball through the virtual world. The performance results show that the goalkeeper can be more effective and more accurate while intercepting a virtual ball than watching the ball ends at the video clip.

Not only VR for athletes but it also the use of VR on physical activity for institutionalized older people [20]. Older people in nursing homes are at risk for decreased physical and cognitive functioning and thus require special attention. The way to reduce that risk is to do a fun physical activity through virtual reality simulation called PhysEx (Physical Excercise with exergames). The study aims to examine the effects of training with PhysEx-based VR on cognitive function, physical performance, depressive symptoms, and fear of falling in institutionalized older people. Besides the VR app for older people, Kiran I et al make VR combined with exergames to produce physical enjoyment activity experiences with high immersive level [21]. Exergames used for driving simulations with added virtual environments in the form of 360 panoramic images from Google Street View.

Another possibility the use of VR is rehabilitation training simulation. Chen et al use VR to train patients with chronic neck pain and kinesiophobia [6]. VR is useful to increase the range of neck motion during rehabilitation training by altering visual feedback. Feedback is taken with Microsoft Kinect device to enhance the visual perception of the user. From that visualization, the user can match his motion during rehabilitation training in a virtual environment.

In this research, we purposed a model of the heart rate monitoring system in a virtual environment. The user will wear a smartwatch and simulate daily exercise simulation through VR application. There is a virtual personal trainer who will provide some physical movement guidance to the user. When the user inside of virtual environment, the user will be able to see the heart rate directly through M2M communication.

III. Methodology

A. Specification Requirement and Analysis

The specification requirements to build this system consist of hardware, software, and network communications modules. The hardware module consists of two parts namely the VR and wearable modules. The VR module consists of an Android smartphone device that has a gyroscope sensor, accelerometer, and proximity (inertial tracking sensor) as well as a VR headset or VR cardboard device. Despite the VR module is a VR mobile device, its possible to replace with VR Desktop devices like Oculus Rift or HTC Vive. The wearable module uses a Samsung Gear 2 smartwatch device with a Tizen operating system which has a heart rate sensor.

The software modules used to develop this system are Unity and Tizen Studio. Unity used to create VR app for daily exercise simulation. Tizen Studio used to create heart rate monitoring application. The network communication module uses the wireless sensor which is attached to smartphone and wearable device.

Daily exercise consists of ordinary muscle stretching movements in hand and foot trough VR simulation. Head or neck motion is not included in this simulation system because it can trigger motion sickness. Motion sickness is a disease who referred to symptoms with easily observed signs. The signs associated with real exposure (physical or visual) or a real motion movement [22].

B. System Design

In general, Fig. 1 depict the proposed VR system. Input on the system in the form of a user's gaze when seeing virtual objects in a virtual environment and a value from the heart rate transmitted through a wearable device. The heart rate data transmission sent by using the internet network. So that wearable device must be connected to a wireless network. After the server receives the requested heart rate data, the data stored in the database. The user will be able to see their heart rate directly through the VR application in a virtual environment.

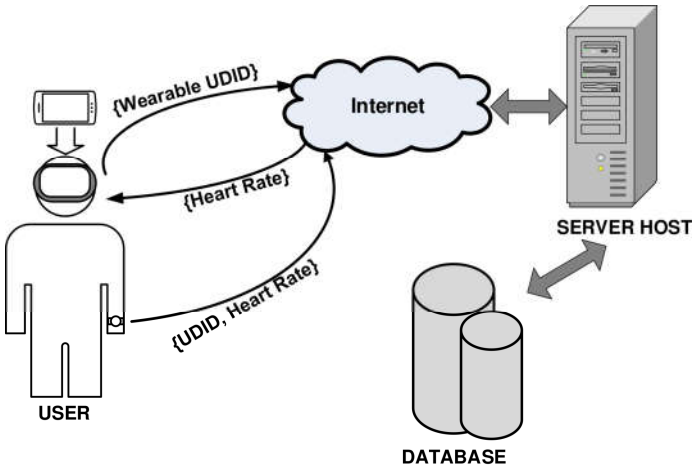


Fig. 1. System architecture.

The first stage to build heart rate monitoring system in a virtual environment is making a heart rate monitoring apps through a wearable device. In this case, we used Samsung Gear 2 as a wearable device. Samsung Gear 2 work in Tizen operating system so the app must build with Tizen programming environment. Tizen application offers two programming environment C++ and Javascript. In this case, we used Javascript as a programming environment. Below are the steps to get the heart rate captured data from a Samsung Gear 2's sensor:

1) Initialization, register health info privilege on Tizen manifest file.
2) Declare a variable to provide the heart rate reference via Tizen API.
3) Create an event-listeners mechanism to start getting the heart rate information.
4) Get heart rate data and put it on application layout.

Heart rate data obtained after waiting for 10 - 15 seconds for system initialization and measuring heart rate for first-time use [9], [10]. Subsequently, for the next seconds, the heart rate would be counting simultaneously. Heart rate value is the number of times a person's heart beats per minute.

The second stage after counting the heart rate is sending the current heart rate data through M2M communication. The selected M2M communications are wireless communication then transmitted over the internet at the server domain address as shown in Fig. 1. The server store heart rate captured temporarily.

URL Request mechanism used to handle the data transmission. Not only heart rate information is sent but also the information of unique device identifier (UDID). UDID required as a primary key. So that heart rate data from a wearable device would not relate to another one. UDID and heart rate stored in the database server as a temporary data storage. Direct communication from wearable devices to smartphones not conducted. It because not only mobile VR can receive the data but also the desktop VR. After the application of heart rate monitoring on the wearable device finished, then create the VR application for daily exercise simulation. In the VR for daily exercise simulation, specify some virtual objects that will stand in the virtual environment (VE). Description of these virtual objects presented in Table I.

TABLE I
VIRTUAL OBJECTS DESCRIPTION

No.	Object Name	Type	Description
1.	Trees	Static	As part of the VE.
2.	Cave	Static	As part of the VE.
3.	Hill	Static	As part of the VE.
4.	Sky	Static	As part of the VE.
5.	Virtual Personal Trainer	Dynamic	As daily exercise trainer.
6.	Instruction text	Dynamic	Text for trainer's instruction.
6.	Heart rate text	Dynamic	Text for heart rate info.

Dynamic objects can update the position both physically and functionally, while static objects are not. After object

specification, the next phase is 3D models creation. The 3D models arranged in predetermined place. There is an animation inside on the virtual personal trainer. The animation is some physical movements that will be simulated. Physical movements divided into two parts: warm-up and simple workout exercise. Warm-up consists of two stretching movements of the feet and hands. The simple workout consists of four movements that call the user to sit on the floor with straight legs parallel and perform some basic physical movement of the feet and hands.

The virtual personal trainer provides guidance with sounds dialog. A visual form of that dialog appears in the text dialog object. This VR exercise app uses Google VR API to display stereoscopic images. Before performing the daily exercise, for first-time use, the user requested to enter a UDID of the wearable device in the text field provided by VR app. UDID obtained from the alert popup in heart rate monitor app shortly after launched. After user inserted the UDID to the server, it will compare with the existing data on the server database. When it found, the virtual personal trainer will start the exercise section. Otherwise, the user will be prompted to enter the correct UDID data.

When a user performs a physical activity instructed by a virtual personal trainer, the user could see his heart rate. Heart rate received by the URL Request mechanism with correct UDID. For each received heart rate value will be stored in the local storage as a data history log.

C. Implementation

Implementation of the user starts by using wearable devices. Then, the user runs a heart rate monitoring app on the wearable device to see the heart rate as in Fig. 2. Furthermore, the user starting the VR app for daily exercise on his mobile phone. This VR app is tested on the Xiaomi Redmi Note 4 as shown in Fig. 3. In this condition, the user will not be able to see the heart rate directly on the wearable device because the user is already inside of virtual environment. The user will not directly perform physical activity but wait for instructions from the virtual personal trainer. The user performs physical activity (warm-up and simple workout exercise) for a total of 6 minutes.



Fig. 2. Heart rate app.

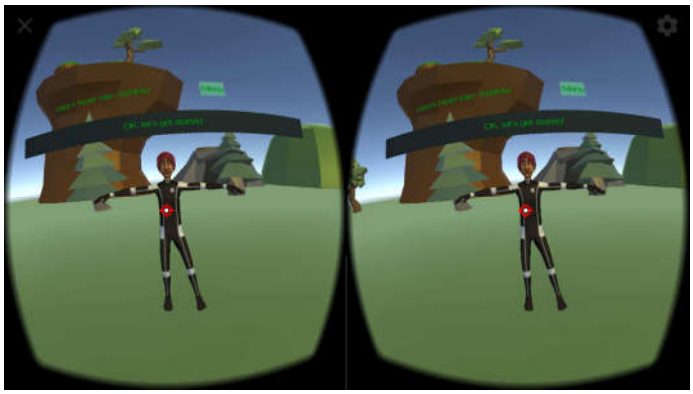


Fig. 3. VR Excercise app.

Fig. 4 demonstrates when a user performs the simple workout exercise in a virtual environment with advising from a virtual personal trainer. In the upper left corner of the personal trainer, there is a user's heart rate that will appear and changes every five seconds. However, the Samsung Gear 2 sensor is still calculated for every single second to take the heart rate. Thus, the user will be able to see the initial heart rate and compare their heart rate while doing physical activity in a virtual environment. There is an early warning system if the user's heart rate in a virtual environment is higher than 130 Beats Per Minute (BPM) or lower than 60 BPM. With these warnings, the user can stop or rest immediately after knowing his weak condition while doing activity in the virtual environment.



Fig. 4. User implementation

IV. Experiments & Result

The user's evaluation performed when the user is finished using the VR app and heart rate monitoring app. The evaluation method is usability testing. Usability testing in the

virtual environment adapted from the System Usability Scale or SUS questionnaire proposed by Lewis et al [23]. There are ten items in SUS questionnaire. We taking only five items which are on item number 1, 9, 3, 4, and 6. This because the usability assessment focuses on the use of heart rate monitoring rather than the entire system in the virtual environment. The whole items are made with the positive worded answer and modified according to the existing system without changing the meaning. The point of scales used with the numbered value from 1 (Strongly disagree) to 5 (Strongly agree).

1) I think that I would like to use this heart rate monitoring frequently while in the virtual environment (Q1).
2) I felt very confident to use this VR with the addition of heart rate monitoring system (Q2).
3) I felt this heart rate monitoring system was easy to use (Q3).
4) I could see my heart rate through this VR app directly (Q4).
5) I see my heart rate consistency while performing any activity in the virtual environment through this VR app (Q5).

This research involved 16 participants which consist of seven male and nine female, with an age range of 18-22 years. The results of five statement items of the questionnaire on all participants presented in Table II. The mean score for each statement from Q1 to Q5 has greater than 3.9. The mean score of the Q1 and Q2 statements are used to assess the satisfaction factor. The mean score of Q4 and Q5 statements are used to assess the effectiveness factor. The efficiency factor only calculated from the mean value of the Q3 statement. The factors which have the highest influence is effectiveness factor with a mean score of the effectiveness of 4.15625 then followed by the satisfaction of 4.09375. The mean score of Efficiency factor has the lowest value of 3.9375. Nevertheless, the mean of overall assessment result taken by participants for all factors is 4.0875. It means participants agree with all statements.

TABLE II
THE SUS QUESTIONNAIRE RESULT

Participant	Q1	Q2	Q3	Q4	Q5
1.	5	5	4	4	2
2.	4	4	4	4	4
3.	5	5	5	5	5
4.	5	5	5	5	5
5.	5	4	4	4	4
6.	4	4	4	5	3
7.	4	3	4	4	4
8.	2	4	4	4	5
9.	2	3	3	4	3
10.	4	4	3	5	4
11.	4	5	3	5	5
12.	4	4	3	4	4
13.	4	3	4	4	4
14.	5	5	4	5	5
15.	4	4	4	4	4
16.	4	4	5	4	4
Mean	**4.0625**	**4.125**	**3.9375**	**4.3125**	**4**
Satisfaction	**4.09375**				
Efficiency	**3.9375**				
Effectiveness	**4.15625**				
Overall	**4.0875**				

Heart rate results stored in the log of history data, so it can be used to determine heart rate values when monitoring heart rates in a virtual environment through this VR app. Fig. 5 shows the heart rate values of 6 participants while performing physical activities in a virtual environment through VR app.

Participants divided into two types: male participants (dashed line) and female participants (solid line). In Fig. 5, depicts the heart rate (BPM) change in every five seconds while performing physical activity (exercise). When the physical activity ends (or is almost over), the heart rate is seen gradually returning to its original value before performing any physical activity.

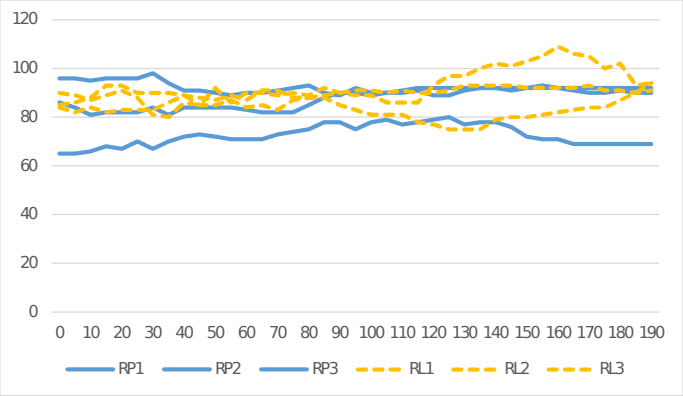


Fig. 5. Heart rate monitored during activity.

V. CONCLUSION

A heart rate monitoring system in a virtual environment is possible to do with the support of wearable devices. In our case studies on VR applications for exercise, it can force participants to perform physical activities. The participants can see his heart rate values changing while performing physical activity. Participants felt that the heart rate monitoring system was satisfied, effective, and efficient measured by an adapted SUS questionnaire.

In future development, it is possible to assess the user's health not only from the heart rate sensor. The potential of embedded sensors in wearable devices allows developers to integrate systems on these devices with virtual reality systems. M2M communication through an internet network (wireless) can be used to integrate the VR system and wearable device system. Besides, it is important to secure the user data usage and user privacy while monitoring users health in a virtual environment.

ACKNOWLEDGMENT

Thanks to the Directorate of Research and Community Service, the Directorate General for Research and Development of the Ministry of Research, Technology and Higher Education

of the Republic of Indonesia who has supported and provided funds for the continuation of this research.

REFERENCES

[1] J. Jerald, The VR Book. *Association for Computing Machinery*, 2015.

[2] K. R. Anderson, M. L. Woodbury, K. Phillips, and L. V. Gauthier, "Virtual reality video games to promote movement recovery in stroke rehabilitation: A guide for clinicians," *Arch. Phys. Med. Rehabil.*, vol. 96, no. 5, pp. 973-976, 2015.

[3] A. Mostafa, W. Ryu, S. Chan, K. Takashima, and G. Kopp, "Designing NeuroSimVR: A Stereoscopic Virtual Reality Spine Surgery Simulator," *Science (80-.).*, pp. 1-20, 2017.

[4] B. Tan, Y. Cai, Y. Zhang, X. Wu, Y. Chen, and B. Yang, "Virtual reality continuum for heritage at Haw Par Villa in Singapore," *in Proceedings of the Symposium on VR Culture and Heritage - VRCAI 16*, 2016, pp. 71-74.

[5] E. Olmos, J. F. Cavalcanti, J.-L. Soler, M. Contero, and M. Alcaiz, "Mobile Virtual Reality: A Promising Technology to Change the Way We Learn and Teach," *in Mobile and Ubiquitous Learning: An International Handbook*, 2018, pp. 95-106.

[6] K. B. Chen et al., "Use of Virtual Reality Feedback for Patients with Chronic Neck Pain and Kinesiophobia," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 8, pp. 1240-1248, 2017.

[7] N. Miyai, "Blood Pressure Response to Heart Rate During Exercise Test and Risk of Future Hypertension," *Hypertension*, vol. 39, no. 3, pp. 761-766, 2002.

[8] S. Genovesi, D. Zaccaria, E. Rossi, M. G. Valsecchi, A. Stella, and M. Stramba-Badiale, "Effects of exercise training on heart rate and QT interval in healthy young individuals: Are there gender differences?," *Europace*, vol. 9, no. 1, pp. 55-60, 2007.

[9] U. NUSSINOVITCH et al., "The Efficiency of 10-Second Resting Heart Rate for the Evaluation of Short-Term Heart Rate Variability Indices," *Pacing Clin. Electrophysiol.*, vol. 34, no. 11, pp. 1498-1502, Nov. 2011.

[10] D. Giles, N. Draper, and W. Neil, Validity of the Polar V800 heart rate monitor to measure RR intervals at rest, *Eur. J. Appl. Physiol*., vol. 116, no. 3, pp. 563-571, Mar. 2016.

[11] V. L. Kalyani and D. Sharma, "IoT: Machine to Machine (M2M), Device to Device (D2D) Internet of Everything (IoE) and Human to Human (H2H): Future of Communication," *J. Manag. Eng. Inf. Technol.*, no. 26, pp. 2394-8124, 2015.

[12] M. B. Almeida and C. G. S. Arajo, "Effects of aerobic training on heart rate," *Rev Bras Med Esporte*, vol. 9, no. 2, pp. 113-120, 2003.

[13] A. Pantelopoulos and N. G. Bourbakis, "A Survey onWearable Sensor-Based Systems for Health Monitoring and Prognosis," *IEEE Trans. Syst. Man, Cybern. Appl. Rev.*, vol. 40, no. 1, pp. 1-12, 2010.

[14] P. Marc-Alexis Remond, "Man vs. Machine: M2M Needs H2H to Become Smarter Collaboration Makes It Possible," 2014. [Online]. Available: https://www.wired.com/insights/2014/09/man-vs-machine/. [Accessed: 12-Apr-2018].

[15] F. El-Amrawy and M. I. Nounou, "Are currently available wearable devices for activity tracking and heart rate monitoring accurate, precise, and medically beneficial?," *Healthc. Inform. Res.*, vol. 21, no. 4, pp. 315-320, 2015.

[16] P. Brey, "Virtual Reality and Computer Simulation ," *Handb. Inf. Comput. ethics*, pp. 361-384, 2008.

[17] P. Halarnkar, S. Shah, H. Shah, H. Shah, and A. Shah, "A Review on Virtual Reality," *IJCSI Int. J. Comput. Sci*. Issues, vol. 9, no. 6, pp. 325-330, 2012.

[18] J. Wang, "Research on application of virtual reality technology in competitive sports," *Procedia Eng.*, vol. 29, pp. 3659-3662, 2012.

[19] N. Vignais, R. Kulpa, S. Brault, D. Presse, and B. Bideau, "Which technology to investigate visual perception in sport: Video vs. virtual reality," *Hum. Mov. Sci.*, vol. 39, pp. 12-26, 2015.

[20] R. S. Monteiro-Junior et al., "Virtual RealityBased Physical Exercise With Exergames (PhysEx) Improves Mental and Physical Health of Institutionalized Older Adults," *J. Am. Med. Dir. Assoc.*, vol. 18, no. 5, p. 454.e1–454.e9, 2017.

[21] K. Ijaz, Y. Wang, N. Ahmadpour, R. A. Calvo, and S. Member, "Physical Activity Enjoyment on an Immersive VR Exergaming Platform," pp. 59-62, 2017.

[22] B. D. Lawson, "Motion Sickness Symptomatology and Origins, *In Handbook of virtual environments: design, implementation, and applications, second edition, edited by Kelly S. Hale and Kay M. Stanney, CRC Press*, 2014," *in Behaviour & Information Technology*, vol. 34, no. 7, 2015, pp. 752-754.

[23] J. R. Lewis and J. Sauro, "The factor structure of the System Usability Scale," vol. 1, pp. 1-10, 2009.

Facial Emotion Recognition based on Viola-Jones Algorithm in the Learning Environment

Kartika Candra Kirana [1], Slamet Wibawanto [2], Heru Wahyu Herwanto [3]
Department of Electrical Engineering,
State University of Malang
Jl Semarang No 5, Malang, Indonesia
kartika.candra.ft@um.ac.id [1], slamet.wibawanto.ft@um.ac.id [2], heru_wh@um.ac.id [3]

***Abstract*—An emotion is a trigger of learning success, so the learning should be adapting to the students' emotions. The most of popular approach is the acquisition of facial-based features. Therefore, we present facial emotion recognition based on the Viola-Jones Algorithm in the learning environment. Basically, the Viola-Jones algorithm is a face detection algorithm. However, we use facial-based features to detect face and recognize emotion, thus we applied rectangular feature and cascading AdaBoost algorithm which are the main concept of the Viola-Jones Algorithm in those both of process. In this study, we compare accuracy, precision, recall, and time-consuming of the Viola-Jones algorithm and our previous methods [1] using 50 UM's learning images in student emotion recognition. The accuracy, precision, recall, and time-consuming of Viola-Jones algorithm reach 0.74, 0.73, 0.76 and 15 seconds per frame, whereas our previous methods [1] reach 0.46, 0.48, 0.52, and 42 seconds per frame. In emotional recognition, we can conclude that the viola jones algorithm is superior to our previous research.**

Keywords—facial recognition, emotion recognition, learning emotion, Viola-Jones algorithm

I. INTRODUCTION

An emotion often leads to a trigger of learning success. The chance of successful learning of an enthusiastic student is greater than a bored student. It caused by the 'enthusiastic' emotion that leads to a willingness to encourages the body and the brain to learn and analyze information [1][2]. Therefore the learning should be adapting to the student's emotion.

There are many approaches to recognize the emotion. The most of popular approach is the acquisition of facial-based features [3]. In the real classrooms, the facial-based features can be captured by CCTV. Whereas in the online learning environment, it can be captured by webcam. The further process is applied after the acquisition is worked for getting the analysis result. One of the most commonly implemented algorithms is the Viola-Jones algorithm.

The Viola-Jones algorithm is real-time processing for face detection. It designed by [4] by applying AdaBoost algorithms for rectangular features classifier in "cascade" stages. Their primitive concept allows background image are discarded quickly. Their face detection proceeded in 15 frames per second (fps). However, it is less responsive to changes in face angle.

Some studies modify the Viola-Jones algorithm to optimize face detection performance. [5] minimize Viola Jones Algorithms into II stages. The result of several frontal cascade face which is tested shows that the convergence of their classifiers better than original Viola-Jones algorithms.

The other optimization is produced by the enhancement of data representation. [6] proposed optimization of data integration to handle the limitation of rectangular features. The rectangular features implement the image integral which caused it less responsive to changes in face angle. Therefore, they insert rotation in the matrix of image integral.

One of these algorithm optimizations is developed through the improvement of hardware specifications. [7] proposed a field programmable gate arrays (FPGA). It is an important building block in video tracking. Image that size of 320x240 pixels validated on the DE2-115 evaluation board. The result optimized the performance of Viola-Jones Algorithms in 25 fps.

In the learning environment, [8] have been applied the combining of Viola-Jones Algorithms for face detection and a neural network for emotion recognition. However, their proposed methods applied in the database which does not present the real expression of student emotion. In addition, they consumed 112.6 until 154.16 seconds of each image for emotion recognition. Their long-term consumption is influenced by the design of neurons in the neural network. However, neural networks reach the high accuracy as shown by [8] which reach the accuracy from 78.54%

[1] implements artificial neural networks and fisherface features to recognize emotions in learning. as generated by [8], they produce an accuracy of 81%. However, they confirm that they apply a complex neural network architecture to produce a system that works accurately but runs slowly. In addition, the fisher face feature also takes a longer load time than the primitive features. Therefore, in this study, we focus on applying primitive features to get faster computing time.

The above studies have been shown that the Viola-Jones algorithm is a robust algorithm for face detection. Moreover, facial-based features are extracted in both of stages consisting of face detection and emotion recognition. These reasons inspire us to combine face detection and emotional recognition using the same classifier which is applying of Viola-Jones

Algorithms. It hoped to prevent a prolonged period of time in facial emotion recognition. Furthermore, we propose Facial Emotion Recognition based on the Viola-Jones Algorithm in the Learning Environment.

II. Emotion on Learning Environment

The basics emotion is divided into the positive emotion and the negative emotion. The basics of emotion is divided into the specified emotion which is consisting of neutral, sadness, happiness, fear, anger, surprise and disgust. The emotion can be captured by facial expression, speech, gestures, and control. [1][2][9]. However, the most of easy acquisition in emotion recognition is the facial-based features [3].

Basic emotion cannot determine the process of learning in detail. Thus the learning expert classifies the learning emotion into a specified emotion. There are two popular learning emotion approaches consisting of (1) Russell's Circumplex approach and (2) the effective and constructive learning-based approach. [10][11]

The approach of Russell's Circumplex classify emotion into (1) satisfaction, (2) hopefulness, (3) boredom, (4) interest, (5) engagement, (6) confusion, (7) frustration, and (8) disappointment. This approach describes emotions in detail. However, there are some ambiguous emotions such as interest and satisfaction. Basically, the satisfaction was obtained when the interested student has finished the learning. However, both emotions are difficult to break down into the features. In addition, not all of these emotions are realizable in class.[1]

The effective and constructive learning-based approach determine emotions more simply. It separated the emotion into four types, namely: (1) interested in a matter of learning, (2) confusion and difficulties in a matter of learning, (3) frustration in learning, and (4) achieve new ideas [11]. However there are some ambiguous features such as confusion/difficulties and frustration [1].

In this research, we simplify the emotion based on the effective learning approach. Thus, we recognize two basic student emotions consisting of 'interest' and bored. 'Interest' presents the positive emotion which is described as being willing to transfer and digest knowledge in the brain. Nevertheless, 'bored' present negative emotion which is described as a state of not willing to transfer and digest knowledge in the brain.

III. Methodology

The proposed method consists of image acquisition, preprocessing, face detection and emotion recognition. In this study, the face detection and the emotion recognition were incorporated using the Viola-Jones Algorithm. Design of the proposed method is shown in Fig. 1.

A. Image Acquisition

In the acquisition phase, we recorded ten student expression in the class using the CANON EOS DSLR camera. Every student is positioned facing the camera. In our data, sunlight can enter the room. However, we do not pre-process light intensity. Files are saved in *'.MOV'* format with a width of 1280 pixels, a height of 720 pixels, and a frame rate of 25 fps. The data sample is shown in Fig. 2.

B. Preprocessing

After we record student emotions in video format, these videos are converted into one frame of face images per 25 seconds iteratively and automatically, thus we get 100 images of student facial expression. The composition of the testing data and training data is 50:50. There were several students recorded in one picture, so that the 100 images contained 200 emotions. The intensity of light disturbs our data. However, we will overcome it in future research.

C. Face Detection and Emotion Recognition using Viola-Jones Algorithms

In this study, we modified the Viola-Jones algorithm for both of process consisting of face detection and emotion recognition. The pseudocode Viola-Jones algorithm is shown in Fig. 3.

Viola-Jones algorithm is a cascaded Adaboost classifier based on image integral. Every sub window shift, the integral image is calculated. The integral image is the summation of the above and left pixel of *x, y*. The representation of the integral image is shown in Figure 4. Integral image can be calculated using Equation 1.

$$w_e = \Sigma w_i^m, \quad (1)$$

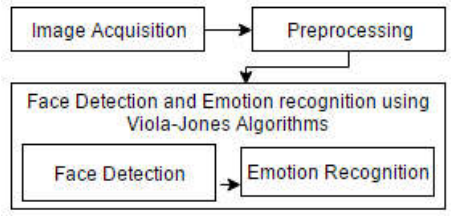


Fig. 1. The Design of Proposed Method



Fig. 2. The Data Sample

```
Algorithm: Viola-Jones algorithms for Face recognition and Emotion Recognition
Input: image
Output: state of emotion
for i ← 1 to shift total do
 for j ← 1 to stages total do
   for k ← to filter total do
    we = integral of subwindows
    update weight_face
    update weight_emotion
    if we < weight_face then
     break for k_loop
    else
     wtot = Σ we
    end_if
   end_for
 end_for
 if wtot < weight_ emotion then
    output = bored
 else
    output = interest
 end_if
end_for
```

Fig. 3. Pseudocode of Viola-Jones Algorithms for Emotion Recognition

where w_e is the weight of sub window's and w_i is the weight of pixel-*i*. In order to select the face area, the image integral is classified using the cascaded AdaBoost classifier.

In a cascading AdaBoost classifier, *M* filters are reapeated at each stage. The weak classifier (k_m) is extracted from the classifier pool. Then, the weight of classifier w_m is set as follows:

$$w_m = ½ \, (ln(1\text{-}e_m) - ln(e_m)) \,, \tag{2}$$

where e_m is normalized of weight which is calculated by:

$$e_m = \frac{W_e}{W}, \tag{3}$$

where *W* is a maximum weight. Then, the threshold w_t is updated by:

$$w_t^{m+1} = \begin{cases} w_t^m e^{\alpha_m} & k_m(x_i) \text{ is a miss} \\ w_t^m e^{-\alpha_m} & \text{otherwise} \end{cases}, \tag{4}$$

The updated threshold w_t on the next iteration ($m+1$) is compared with the weight of the sub-windows w_e. If the weight of sub window is less than the threshold, the subwindows are rejected as the face. Whereas if the weight of sub window is more than the threshold, the sub window is processed in emotion recognition.

The procedure of emotion recognition is similar to face detection which is consisting of calculate the integral image and the classify using the cascaded AdaBoost. However, the integral image at this stage is the sum of the filtered faces in the previous stage which is shown as follows:

$$w_{tot} = \Sigma w_e^m, \tag{5}$$

where w_{tot} is the integral image of the filtered faces-w_e^m. In this stage, the threshold is updated on each iteration ($m+1$), by applying Equation (4). The updated threshold is compared with the weight of filtered faces. If the weight of filtered faces is less than the threshold, the filtered faces are set as 'interest' expression. Whereas if the weight of filtered faces is more than the threshold, the filtered faces are set as 'bored' expression. In this study, we assume that interest expressions have a greater value than the threshold because some indicators such as smile increase the value of the texture feature that allowing an increased integral value.

D. Evaluation

We set a series of evaluation for getting the performance of Viola-Jones algorithms for facial emotion recognition. The emotions ground truth is analyzed manually by an expert. The estimated results are compared with ground truth to obtain the accuracy, precision, recall, and time requirements. We calculate time requirements in second per frame by averaging the rendering time in the test data. Meanwhile, the accuracy, precision, and recall are calculated as follows:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

$$precision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

where *TP* is the total of the interest expression which classified correctly, *TN* is the total of the bored expression which classified correctly, *FN* is the total of interest expression which classified as bored expression, and *FP* is the total of bored expression which classified as interest expression. This evaluation was applied to 100 images we recorded in learning at the State University of Malang. However, there were several students we recorded together in one picture, so 100 images contained 200 emotions. If there are two faces in one image, we calculate them separately. For example, if there are interest and bored in one image, we count it as two objects. Interest object has the possibility of *TP* and *FN* values, while bored object has the possibility of *TN* and *FP* values.

IV. Result and Discussion

Series of experiments are conducted to evaluate the number of the stage for getting the best performance of Viola-Jones Algorithms. The result of stage evaluation is shown in Table I. Based on Table I, the use of a little stage cause the system cannot distinguish the emotions. However, too many stages consume a lot of time, but do not raise accuracy, precision, and recall significantly, so we use three stages. Stage I is used to reject unface area, stage II is used to update the weight of face, and stage III is used to estimate the emotion using the updated weight.

We also evaluate the number of rectangle feature for getting the best performance of Viola-Jones Algorithms. The result of stage evaluation is shown in Table II. The result shows that the use of little features causes the system cannot distinguish the emotion. However, many features consume a lot of time, but do not raise accuracy, precision, and recall significantly, thus in this study, we use nine-rectangle features.

TABLE I. THE EVALUATION OF THE NUMBER STAGES

The Number of Stages	Result			
	accuracy	*precision*	*recall*	*Time (second per frame)*
2	0.60	0.65	0.65	9.6
3[a]	**0.74**	**0.73**	**0.76**	**15**
4	0.74	0.73	0.76	19.2
5	0.76	0.76	0.76	24

[a] The value we chose for further evaluation

TABLE II. THE EVALUATION OF THE NUMBER RECTANGULAR FEATURES

The Number of Rectangular Features	Result			
	accuracy	*precision*	*recall*	*Time (second per frame)*
4	0.66	0.65	0.60	6
9[b]	**0.74**	**0.73**	**0.76**	**15**
16	0.74	0.76	0.76	27

[b] The value we chose for further evaluation

TABLE III. THE COMPARISON OF EMOTION RECOGNITION

The Algorithm	Result							
	TP	*TN*	*FP*	*FN*	*accuracy*	*precision*	*recall*	*Time*[c]
Viola-Jones	38	36	14	12	0.74	0.73	0.76	15
Viola-Jones + Neural Network [8]	40	36	14	10	0.76	0.74	0.8	38
Neural Network [1]	26	22	28	24	0.48	0.48	0.52	42

[c] Second per frame

After getting the best number of stages and rectangular features, we evaluated the accuracy, precision, and recall of the system using 50 data of the University of Malang's learning dataset which contains 100 emotions. The result of accuracy, precision and recall are shown in Table III.

In Table III, we compare our proposed algorithm with [8] and [1]. In this study, we use the Viola-Jones algorithm as a face detection algorithm and an emotion classifier. Meanwhile [8] uses the Viola-Jones algorithm as a face detection algorithm and a neural network algorithm as an emotion classifier. Whereas [1] uses neural network algorithms as face detection algorithm and emotion classifiers. The results show that our proposed algorithm achieves accuracy, precision, recall, and time up to 0.74, 0.73, 0.76 and 15 respectively. Whereas [8] algorithms achieves 0.78, 0.74, 0.80, 38 and [1] algorithm reached 0.48, 0.48, 0.52, and 42. The three sets of results show that the best accuracy, precision, and recall algorithms are a combination of neural networks and the Viola-Jones algorithm [8], whereas the fastest algorithms are Viola-Jones algorithms.

The Viola-Jones algorithm not only is applied in our proposed algorithm but also [8]'s algorithm, thus the number of detected faces is similar. However, both methods suggest that the Viola-Jones algorithm detects faces facing the camera well, but does not detect a face that has a large face tilt. This is shown in Figure 4(a) and Figure 4(c).

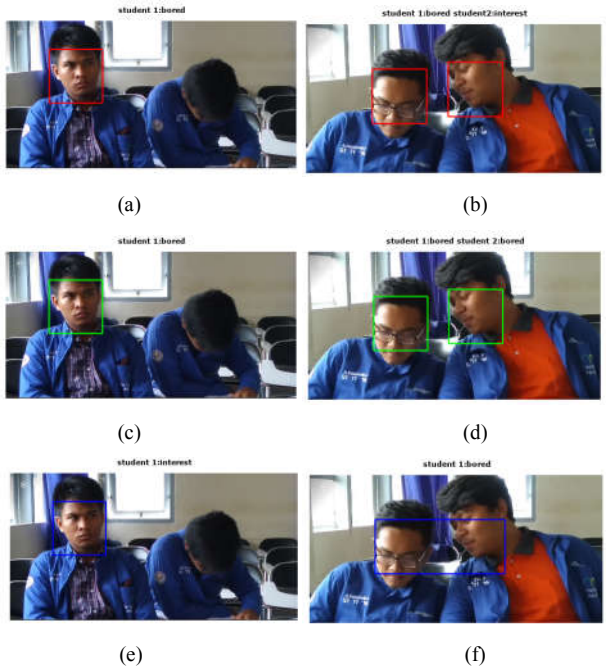


Fig. 4. The Result Comparison of Proposed Method, Viola-Jones+Neural Network Algorithm [8], and Neural Network Algorithm [1] (a),(b) Result of Proposed Method (c)(d) Result of Viola-Jones+Neural Network Algorithm [8] (e)(f) Result of Neural Network Algorithm [1]

Neural network in [8]'s algorithm improves emotion recognition shown in Figure 4(d). The neural network detects both students' emotions appropriately which is bored emotions. This is demonstrated by the two students who did not focus on the presentation slides and discuss other topics.

Based on Figure 4(b), our method and [8]'s algorithm wrongly detects the emotion of one of the students. This is due to a student smile that increases the image's integral on multiple rectangular blocks. However, neural networks do not increase accuracy significantly, but increase computing time doubled from our method.

Meantime, the implementation of the algorithm [1] in our database did not work properly. [1]'s Algorithm that only detects the a* and b* based faces produced a high face detection error when the background color and foreground color is similar. The face detection errors decrease the accuracy of emotion recognition which is shown in Fig. 4(e) and Fig. 4(f). However, the comparisons of Figures 4(a), 4(c), and 4(e) show that if the face area is detected correctly, the fisher face feature is more precise in detecting emotion than the rectangular feature.

V. CONCLUSION

Series of experiments are conducted to evaluate the performance of Viola-Jones algorithms. We also compared with [8] and [1] as comparison algorithms. The accuracy, precision, recall, and time-consuming of Viola-Jones algorithm reach 0.74, 0.73, 0.76 and 15 seconds per frame, while [1] reach 0.46, 0.48, 0.52, and 42 seconds per frame. In emotional

recognition, we can conclude that the viola jones algorithm is superior to [1] as our previous research. However, the accuracy, precision, and recall of Viola-Jones algorithms were lower than [8]. Moreover, the increase [8] is not significant and consumes twice as much time as our method, thus, we can conclude that the viola jones algorithm is a rapid algorithm in the facial emotion recognition. However, in further research, the improvement of accuracy, precision, and recall of this algorithm can be done by improvising the features used which remain attentive to the consumption of time spent. Trials on larger data and improvements to face detection errors that do not face forward also need to be done.

ACKNOWLEDGMENT

This work fully is funded using PNBP of the State University of Malang. We also express our gratitude to the 2016 Informatics Engineering students and the computer vision laboratory team who have assisted in retrieving the database

REFERENCES

[1] S. Wibawanto and K.C.Kirana, "Recognition of Student Emotion based on Matrix-1 Median Fisher's Face and Backpropagation Algorithm," 2017 International Conference on Electrical Engineering and Informatics, pp. 117-122. October 2017.

[2] L. Darling, S. Orcutt, K. Strobel, E. Kirsch, I. Lit, and D. Martin., The Learning Classroom: Emotion and Learning, California: Stanford University School of Education, 2010, pp. 90-104.

[3] P. Tarnowski, M. Kołodziej, A. Majkowski, and R. J. Rak, "Emotion Recognition Using Facial Expressions," International Conference on Computational Science, pp. 1175–1184, June 2017.

[4] P. Viola and M.J. Jones, "Robust Real Time Face Detection" in International Journal of Computer Vision. vol. 57 (2004), pp.137-154.

[5] J.B.Wen, Y.S. Xiong, and S.L Wang, "A Novel Two-Stage Weak Classifier Selection Approach for Adaptive Boosting for Cascade Face Detector" Neurocomputing, vol. 116 (2013), pp. 122-135.

[6] M.V. Alyushin, V.M.Alyushin, and L.V. Kolobashkina, "Optimization of the Data Representation Integrated Form in the Viola-Jones Algorithm for a Person's Face Search," 8th Annual International Conference on Biologically Inspired Cognitive Architectures, Procedia Computer Science, vol. 123 (2018), pp. 18–23.

[7] P. Irgens, C. Bader, T. Lé, D. Saxena, C. Ababei, "An Efficient And Cost Effective FPGA based Implementation of The Viola-Jones Face Detection Algorithm," HardwareX Elesevier, vol.1 (2017), pp. 65-77.

[8] D. Yang, A. Alsadoon, P.W.C. Prasad, A. K. Singh, and A. Elchouemi, "An Emotion Recognition Model Based on Facial Recognition in Virtual Learning Environment," 6th International Conference on Smart Computing and Communications, Procedia Computer Science, Vol. 125 (2018), pp. 2–10.

[9] L. Shen, M. Wang, and R. Shen, "Affective e-Learning: Using Emotional Data to Improve Learning in Pervasive Learning Environment," Educational Technology & Society. vol 12, no. 2, 2009, pp. 176-189.

[10] A.R Faria, A. Almeida, C. Martins, R. Gonçalves, J. Martins, F. Branco, "A global perspective on an emotional learning model proposal," Telematics and Informatics, 2016, in press.

[11] P. Vela, P.A. Vela, R.J. Jensen, Emotions, Technology, Design, and Learning: Robots, Emotions, and Learning, Elesevier Inc, 2016, pp.183-197.

Handwriting Recognition using Eccentricity and Metric Feature Extraction based on K-Nearest Neighbors

Eko Hari Rachmawanto, Galang Rambu Anarqi, De Rosal Ignatius Moses Setiadi, Christy Atika Sari
Department of Informatics Engineering, Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
Email: atika.sari@dsn.dinus.ac.id, ganarqiz@gmail.com, moses@dsn.dinus.ac.id, eko.hari@dsn.dinus.ac.id

***Abstract*— The process of handwriting recognition aims to digitize handwriting without having to retype handwriting. This study proposes the extraction of Eccentricity and roundness features based on K-Nearest Neighbors (KNN) to recognize handwriting. To maximize the process of recognition is done several stages of preprocessing stages such as thresholding, noise filtering, and cropping stage to take every character on handwriting image. Each image that has been through the preprocessing process is further extracted features. Eccentricity and metric are the features of the extracted form of each image to be recognized. After obtained the two features are then carried out the training process to group the feature form on each type of letter. Next, KNN is used to classify each letter character. Based on the results of testing the proposed method for recognition of handwriting characters obtained an accuracy of 85.38%.**

Keywords— Handwriting Recognition, Eccentricity, Roundness, Classification, KNN

I. INTRODUCTION

The rapid development in technology is one of the factors that impact on people's survival. The technological revolution and the digital world change the way people work, how to think, and how they work [1]. Automation technology created to facilitate human work is inseparable from the development of techniques and algorithms applied to the field of computer vision, digital image processing, computer graphics, as well as in the field of artificial intelligence. In the development in the field of digital image processing, for example, various techniques developed, both as image processing and for purposes or specific purposes. Working in this modern era requires something that is practical, fast, easy, and accurate, without having to do a long job, for example in terms of raw data processing. Handwriting data must be retyped into digital text to be processed by the computer, it is time-consuming and ineffective [2]. Therefore, this study aims to recognize handwritten characters.

Handwriting has a different pattern or characteristic, so it needs to do feature extraction first. The extraction required to recognize the pattern of each handwriting character is the extraction of the characteristic of the form [3] [4]. However, prior to extracting the characteristic features of each handwriting image, it is necessary to prepare the preprocessing first. The scanned image that has been scanned and then done segmentation process to separate the region of each character contained in a handwritten image that is with thresholding, then eliminating the noise on the image, then cropping on every area of the writing character [5] [6]. Every image that passes through the preprocessing then extracted its features. This is done to distinguish between one character and character. Features of eccentricity and roundness are the features proposed in this study. The value of roundness is obtained by extracting the characteristic size of each writing image first, ie by finding the value of the area and its perimeterEccentricity value close to 1 means the object approaches the shape of the line, and if the eccentricity value approaches the value 0 close to the round shape. This is the opposite of the value of roundness [7]. With two features of the form then it can be done classification of each handwriting character. The algorithm proposed in this research is K-Nearest Neighbors (KNN). The algorithm is chosen to test whether the KNN algorithm is capable of providing good classification results in handwriting recognition [8].

II. PROPOSED METHOD

This research proposes a recognition method that consists of several stages of preprocessing before extracting features and classifying. The preprocessing process consists of several stages, namely, thresholding, noise removal, and cropping. This process will produce a dataset that will be divided into two data testing and training data. Furthermore, extraction and classification are performed. For more details, the steps of the proposed method can see Figure 1.

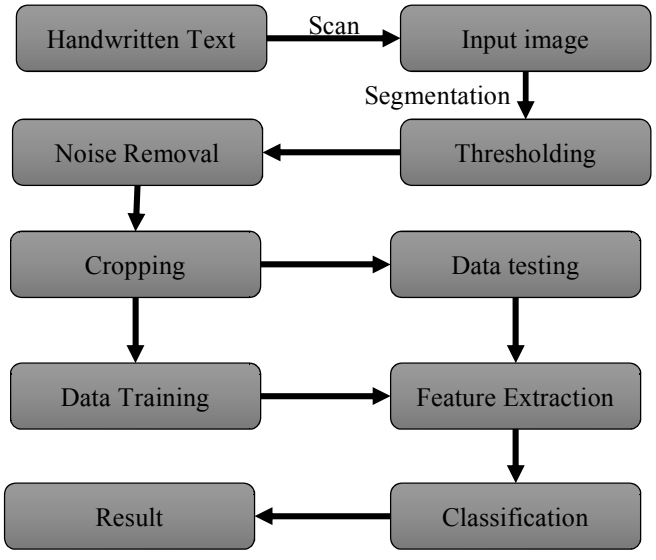


Fig. 1. Stages in the proposed method of recognition

A. Thresholding

Thresholding is the basic transformation to produce a binary image of a gray or color image by setting the pixel value of the image to 1 or 0 depending on it being above or below the threshold value. Thresholding is usually used to separate or segment an area or object in an image based on its

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

pixel value [9] [10]. The threshold pseudocode can be write-in bellow.

```
for each pixel Img(m,n)
    if Img(m,n) > T
        Img(m,n) = 1
    else
        Img(m,n) = 0
    end
end
```



Fig. 2. Threshold result on the image of the character of letter A

B. Noise Removal

Noise seen after the thresholding process must be removed as it may damage the image quality. The most commonly used noise removal methods such as Median Filter. The median filter is also good for maintaining detailed frequency and sharpness of the image edge and also removes noise salt and pepper, this filter is also better than the mean filtering operation [9]. Figure 3 is a sample of a filtering process to eliminate image noise with median filtering.



Fig. 3. Median Filter Result on Image of the character of letter A

C. Cropping

The next step is to look for the region of the image. A region is made by considering the size of a scalar based on its geometric properties. Inside the region, there are areas, perimeters, centroid, and eccentricity. The area is the number of pixels - the pixel of an object whose unit is a pixel that can reflect the size of a real object. The perimeter is the border length of the object. The size of the perimeter is measured by counting the number of pixels from the boundary of the object. the centroid is the coordinate of the midpoint of an object [2]. Eccentricity is the value of the comparison between the distance of a small foci ellipse with foci ellipses major of an object. Eccentricity has a value between 0 to 1. Objects in the form of a straight line / elongate, the value of near eccentricity 1, and objects close to the circle, the value of eccentricity is close to 0. After the region of each character object is obtained, then take the image of each character and put it in the training data folder as well as testing. The result of the cropping process is shown in figure 5.

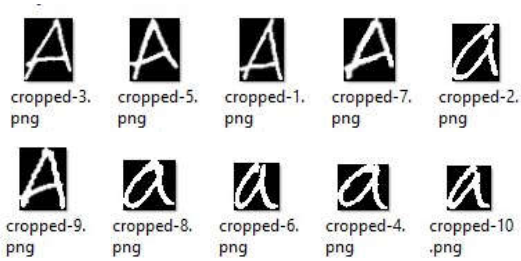


Fig. 4. Cropping results on the character image of the letter A

D. Feature Extraction

The feature extraction process is a process done to get the feature of the image to be classified. In this paper feature extraction performed is the feature extraction form. In feature extraction, there are 2 traits that are used to distinguish 26 types of writing character class ie eccentricity and metric. Eccentricity is obtained based on the value of comparison between the distance of the foci ellipse minor with the foci ellipse major of an object. So it can be known how close the object is close to the form of a circle. Eccentricity is formulated by formula (1).

$$e = \sqrt{1 - \frac{b^2}{a^2}} \quad (1)$$

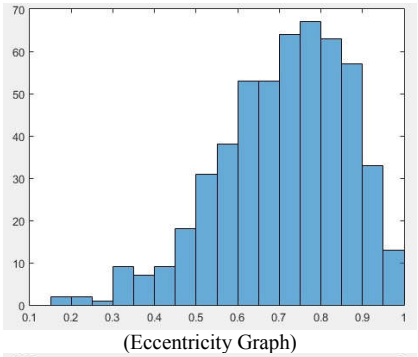
Where :
$e =$ eccentricity
$b =$ minor axis
$a =$ mayor axis

The metric parameter is the ratio between the area and the perimeter of the object. Object that metric close to 0 then the shape of the object close to the form of a long line, while the object metric approaching the number 1 then the shape of the object close to the round shape. Metric calculations are illustrated in formula (2).

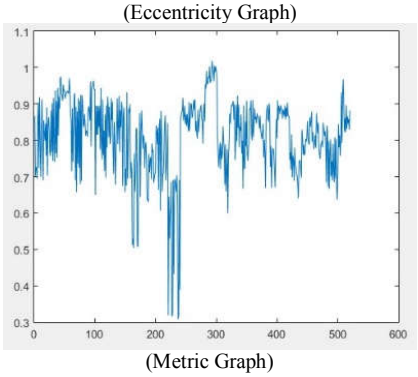
$$M = \frac{4\pi A}{P^2} \quad (2)$$

Where :
$M =$ Roundness
$A =$ Area
$P =$ Perimeter

To create training data, several data types are used on each handwriting character type obtained from the preprocessing performed previously. The data used as training data and then extracted the eccentricity and metric features then grouped by each class character letter. Figure 5 is a sample graph of metrics and eccentricity in training data.



(Eccentricity Graph)



(Metric Graph)

Fig. 5. Metric and Eccentricity Graphics on Data Training

E. Classification

KNN algorithm is one of the methods to classify the object based on training data that is closest to the object. Training data is projected onto the dimensional space where each dimension presents the features of the data. KNN uses a supervised algorithm [11]. The purpose of the KNN algorithm is to classify new objects based on attributes and training data. Where the results of the new test samples are classified by the majority of the categories in KNN. In the classification process, this algorithm does not use any model to be matched and based solely on memory. The KNN algorithm uses the adjacent classification as the predicted value of the new test sample. The distance formula used in this algorithm is Euclidean distance. Ecludian distance is the most commonly calculated distance formula used in numerical data [12]. Distance calculation with Ecludian distance is done by formula (3).

$$D = \sqrt{\sum_{i=1}^{P} (X_{2i} - X_{1i})^2} \quad (3)$$

Where :
$D = distance$
$X_{1i} = Data\ training$
$X_{2i} = Data\ testing$
$i =$ Data $Variabel$
$P =$ D$ata\ Dimension$

In the training phase, this algorithm only stores the feature vectors and the classification of the training data. In the classification phase, the same features are calculated for test data whose classification is not yet known. This algorithm determines the distance value in the test data test with the data testing based on the smallest value of the nearest neighbor (K) value than predicted to be included in the most classification of the points. The best K value in KNN classification depends on the data. However, the higher the K value will reduce the noise effect on the classification but make the limits of each classification become more blurred and reduce its accuracy.

III. Experiment Results and Analysis

Implementation and testing in this paper are done to know the accuracy of handwriting recognition results by classification method using K-Nearest Neighbors algorithm. Testing is divided into 2 phases namely preprocessing and classification.

A. Pre Processing

Handwriting image data obtained by using a scanner then made the conversion from RGB image to grayscale. After the conversion to grayscale is done, next step is image segmentation. The process of segmentation begins with thresholding method. The threshold value is obtained by calculating the mean of all pixels of the grayscale image divided by the total number of pixels. The average pixel yield obtained is then used to calculate the variance value of 1 - L-1. The optimal indigo threshold is determined from the minimum variance value. The threshold value obtained is used to change the intensity of each pixel in the grayscale image. The intensity of the image that is below the threshold value will be changed to a value of 0 and the intensity of the image above the threshold value will be changed to the value 1. After the segmentation process, filtering is done to eliminate noise in the image of segmentation result using a median filter.

The next step is to search for the region of every letter character that is in the image. This is done to retrieve and cropping of any letter character. In the region there are areas, perimeter, centroid, and eccentricity After the region of each character object written obtained, then take the image of each character to be able to generate the image of training image and image data used for the classification stage.

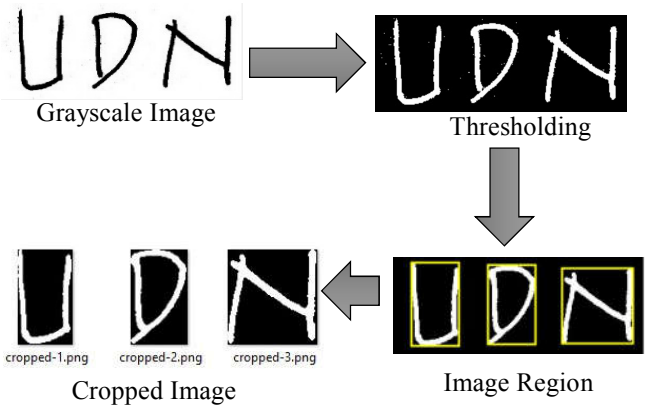


Fig. 6. Preprocessing Stage

However, in the case of taking the region on the characters of the letters (i) and (j) small cannot be done intact because there is a small area separated in the form of a point that is on the letter. The regions of the letters (i) and (j) are shown in Figure 6.



Fig. 7. Region image on character i and j

The result of each character image generated in the pre-processing process is then used as training data or test data. Image data ready to be used as training data or test data, each feature of the form on the letters taken. In the process of extraction of form, there are two traits that are used to distinguish 26 types of character class writing that is eccentricity and metric. Figure 8 is the result of grouping eccentricity and metrics in training data and test data.

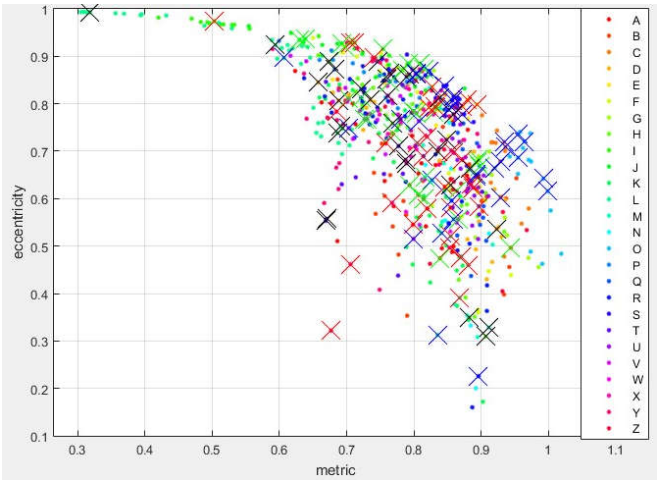


Fig. 8. Metric and eccentricity charts from training data and test data

A point image is a group of letters that feature on the training data and cross-image is a feature group of letters of test data.

B. Classification

In this paper, the training was conducted on 26 types of characters of letters A-Z with the amount of data 20 images in each character so that the total data train as much as 520 data. Total test data is given data as much as 25% of training data, that is as much as 130 data with each character data letters amounted to 5 image images. The images examined vary between handwriting characters of capital letters and lowercase characters of varying sizes. Classification Results is shown in Table I. From the table I can be obtained the total data true amount of 111 and total data wrong number of 19 with the total amount of test data as much as 130. From the above table can know True Positive (TP) and True Negative (TN) from each of the result of classification of letters. The classification results can be explained further by using Confusion Matrix which can provide information on accuracy, recall, and precision. Figure I is respected is the confusion matrix image of the classification with K = 3

TABLE I. RESULTS CLASSIFICATION WITH K = 3, ON LETTERS

Character	Original Class	Results Classification With K = 3, On Letters					True Positive Results	True Negative Results
		1	*2*	*3*	*4*	*5*		
A	1	1	1	1	1	1	5	0
B	2	2	2	2	2	2	5	0
C	3	3	3	3	3	3	5	0
D	4	4	4	4	4	4	5	0
E	5	5	5	5	5	5	5	0
F	6	6	6	6	6	6	5	0
G	7	7	7	7	7	7	5	0
H	8	8	8	8	8	8	5	0
I	9	9	9	9	9	9	5	0
J	10	10	10	10	10	10	5	0
K	11	11	11	11	11	11	5	0
L	12	12	12	12	12	12	5	0
M	13	13	13	13	13	13	5	0
N	14	14	14	14	14	14	5	0
O	15	15	15	15	15	15	5	0
P	16	16	6	16	16	17	3	2
Q	17	17	10	17	14	10	2	3
R	18	11	17	18	16	18	2	3
S	19	19	19	19	19	19	5	0
T	20	20	21	7	20	20	3	2
U	21	21	21	21	21	21	5	0
V	22	22	22	22	22	22	5	0
W	23	23	14	14	14	23	2	3
X	24	24	24	24	24	18	4	1
Y	25	10	25	7	25	2	2	3
Z	26	26	26	2	26	11	3	2
Sum							111	19

Fig. 9. confusion matrix of classification with K = 3

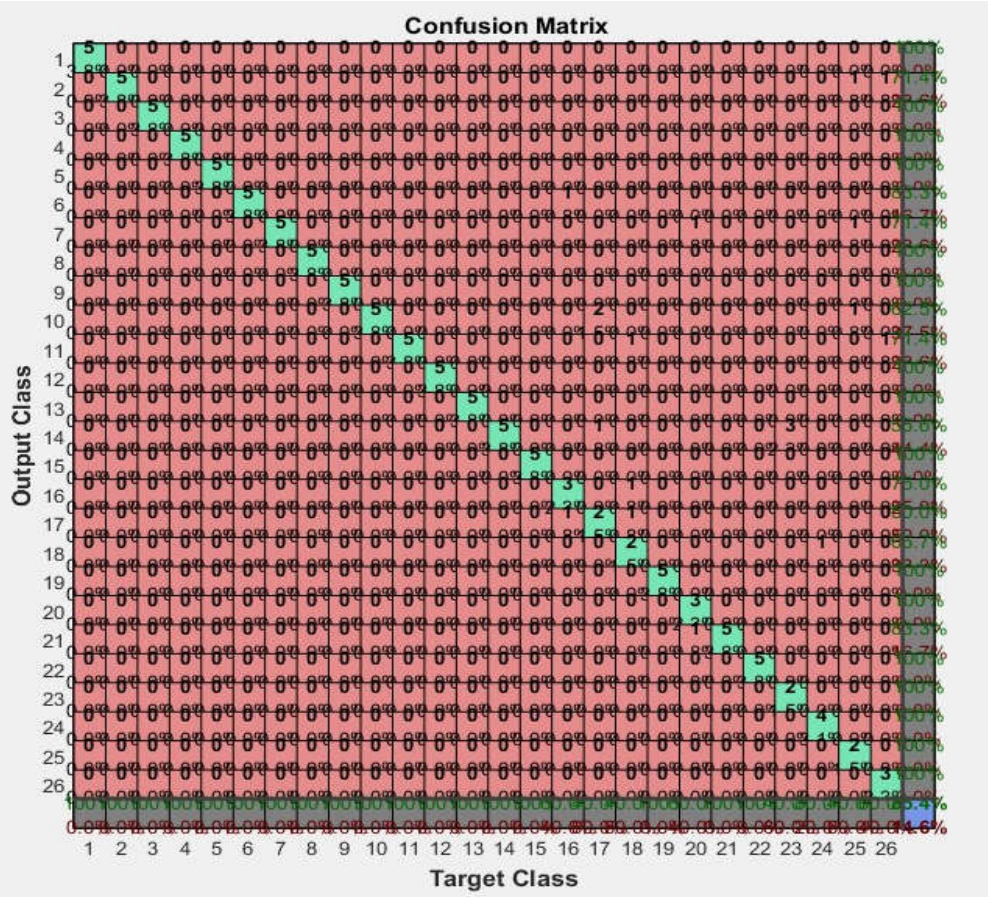


Fig. 10. confusion matrix of classification with K = 3

From the results of the classification confusion matrix using K = 3, accuracy, precision, and recall can be calculated based has been obtained in Table 1.

1) Accuracy

$$Accuracy = \frac{total\ true\ of\ test\ result}{total\ of\ test\ data} \times 100\%$$

$$\rightarrow \frac{111}{130} \times 100\% = 85.38\ \%$$

2) Error Rate

$$Error\ Rate = \frac{total\ false\ of\ test\ result}{total\ of\ test\ data} \times 100\%$$

$$\rightarrow \frac{19}{130} \times 100\% = 14.61\%$$

3) Recall

$$Recall_{A\ldots Z} = \frac{TP_{A\ldots Z}}{amount\ of\ data_{A\ldots Z}}$$

4) Precision

$$precision_{A\ldots Z} = \frac{TP_{A\ldots Z}}{System\ Prediction_{A\ldots Z}}$$

Recall and precision calculation results can be seen in table II below

TABLE II. THE RESULT FROM CLASSIFICATION USED K = 3

Character	Recall	Precision
A	1	1
B	1	0.7143
C	1	1
D	1	1
E	1	1
F	1	0.8333
G	1	0.7143
H	1	1
I	1	1
J	1	0.625
K	1	0.7143
L	1	1
M	1	1
N	1	0.5556
O	1	1
P	0.6	0.75
Q	0.4	0.5
R	0.4	0.6667
S	1	1
T	0.6	1
U	1	0.8333
V	1	1
W	0.4	1

Character	Recall	Precision
X	0.8	1
Y	0.4	1
Z	0.6	1

From the experiments conducted with total training data as much as 520 data handwriting characters and test data as much as 25% of the training data amounts to 130 character data - each character 5 pieces of the image. The result of the accuracy of handwriting character recognition classification reaches 85.38% accuracy. This shows that the classification of handwritten letters using the KNN algorithm is quite accurate.

IV. CONCLUSIONS

The conclusions of the preprocessing results, the processing results were able to take each character of different sizes and patterns on the handwriting image. Median filters are effectively used to remove noise in the image. In the process of searching the region of lower-case characters, the character of the writing that has a separate area eg letter (i) and (j) cannot be performed instantly because there is a small separate area of the point on the letter. Classification using KNN algorithm is relatively good, it is indicated by the result of accuracy which can reach 85.38%. This suggests that the KNN algorithm is quite effective for the classification of handwriting character recognition. This is proportional to the amount of training data that is greater than the test data.

REFERENCES

[1] D. Chaithra dan K. Indira, "Handwritten online character recognition for single stroke Kannada characters," in *IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, Bangalore, 2017.

[2] M. Sepahvand, F. Abdali-Mohammadi dan F. Mardukhi, "Evolutionary Metric-Learning-Based Recognition Algorithm for Online Isolated Persian/Arabic Characters, Reconstructed Using Inertial Pen Signals," *IEEE Transactions on Cybernetics,* vol. 47, no. 9, pp. 2872 - 2884, 2017.

[3] Y. Guerbai, Y. Chibani dan B. Hadjadji, "Handwriting gender recognition system based on the one-class support vector machines," in *International Conference on Image Processing Theory, Tools and Applications (IPTA)*, Montreal, 2017.

[4] C. Zanchettin, B. L. D. Bezerra dan W. W. Azevedo, "A KNN-SVM hybrid model for cursive handwriting recognition," in *International Joint Conference on Neural Networks (IJCNN)*, Brisbane, 2012.

[5] P. R. Paneri, R. Narang dan M. M. Goswami, "Offline handwritten Gujarati word recognition," in *International Conference on Image Information Processing (ICIIP)*, Shimla, 2017.

[6] L. Hiryanto, A. R. Yohannis dan T. Handhayani, "Hand signature and handwriting recognition as identification of the writer using gray level co-occurrence matrix and bootstrap," in *Intelligent Systems Conference (IntelliSys)*, London, 2017.

[7] P. M. Kamble dan R. S. Hegadi, "Geometrical Features Extraction and KNN Based Classification of Handwritten Marathi Characters," in *World Congress on Computing and Communication Technologies (WCCCT)*, Tiruchirappalli, 2017.

[8] T. Makkar, Y. Kumar, A. K. Dubey, Á. Rocha dan A. Goyal, "Analogizing time complexity of KNN and CNN in recognizing handwritten digits," in *International Conference on Image Information Processing (ICIIP)*, Shimla, 2017.

[9] C. Solomon, Fundamentals of Digital Image Processing, 1st Edition, Canterbury, UK: John Wiley & Sons, Ltd, 2011.

[10] S. N dan V. S, "Image Segmentation By Using Thresholding Techniques For Medical Images," *Computer Science & Engineering: An International Journal (CSEIJ),* vol. 6, no. 1, pp. 1-13, 2016.

[11] O. R. Indriani, E. J. Kusuma, C. A. Sari, E. H. Rachmawanto dan D. R. I. M. Setiadi, "Tomatoes Classification Using K-NN Based on GLCM and HSV Color Space," in *International Conference on Innovative and Creative Information Technology (ICITECH)*, Salatiga, 2017.

[12] T. Sutojo, P. S. Tirajani, D. R. I. M. Setiadi, C. A. Sari dan E. H. Rachmawanto, "CBIR for Classification of Cow Types using GLCM and Color Features Extraction," in *International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Yogyakarta, 2017.

Dinus Intelligent Assistance (DINA) Chatbot for University Admission Services

Heru Agus Santoso
Dept. Informatics Engineering
Universitas Dian Nuswantoro
Semarang, Indonesia
heru.agus.santoso@dsn.dinus.ac.id

Galuh Wilujeng Saraswati
Dept. Informatics Engineering
Universitas Dian Nuswantoro
Semarang, Indonesia
galuhwilujengs@dsn.dinus.ac.id

Muhammad Syaifur Rohman
Dept. Informatics Engineering
Universitas Dian Nuswantoro
Semarang, Indonesia
syaifur@dsn.dinus.ac.id

Nurul Anisa Sri Winarsih
Dept.Informatics Engineering
Universitas Dian Nuswantoro
Semarang,Indonesia
nurulanisasw@dsn.dinus.ac.id

Septian Enggar Sukmana
Dept.Informatics Engineering
Universitas Dian Nuswantoro
Semarang, Indonesia
septian.enggar@dsn.dinus.ac.id

Adhitya Nugraha
Dept.Informatics Engineering
Universitas Dian Nuswantoro
Semarang, Indonesia
adhitya@dsn.dinus.ac.id

Edy Mulyanto
Dept.Informatics Engineering
Universitas Dian Nuswantoro
Semarang, Indonesia
edymulyanto@dsn.dinus.ac.id

Supriadi Rustad
Dept.Informatics Engineering
Universitas Dian Nuswantoro
Semarang, Indonesia
srustad@dsn.dinus.ac.id

Fahri Firdausillah
Dept.Informatics Engineering
Universitas Dian Nuswantoro
Semarang, Indonesia
fahri@dsn.dinus.ac.id

Nowadays, universities offer most of their services using corporate website. In higher education services including admission services, a university needs to always provide excellent service to ensure student candidate satisfaction. To obtain student candidate satisfaction apart from the quality of education must also be accompanied by providing consultation services and information to them. This paper proposes the development of Chatbot which acts as a conversation agent that can play a role of as student candidate service. This Chatbot is called Dinus Intelligent Assistance (DINA). DINA uses knowledge based as a center for machine learning approach. The pattern extracted from the knowledge based can be used to provide responses to the user. The source of knowledge based is taken from Universitas Dian Nuswantoro (UDINUS) guest book. It contains of questions and answers about UDINUS admission services. Testing of this system is done by entering questions. From 166 intents, the author tested it using ten random sample questions. Among them, it got eight tested questions answered correctly. Therefore, by using this study we can develop further intelligent Chatbots to help student candidates find the information they need without waiting for the admission staffs's answer.

Keywords—Chatbot, ontology, Neural network, NLP

I. INTRODUCTION

Nowadays, most of universities use a corporate website to offer services to the student candidates. In the field of higher education services, a university needs to always provide excellent service to ensure student candidate satisfaction [1]. To obtain student candidate satisfaction apart from the quality of education, a university must also be accompanied by providing consultation services and information to student candidates. The use of web-based student candidate service system and social media is one of the facilities that are given to student candidates to meet the needs of information. Along with the increase in the world of web services, many of the latest innovations in development focus on developing websites that pay great attention to user convenience [2]. In website itself, there are various types of student candidate service available such as phone service and live chat support. All available supports aim to communicate between peoples, therefore it needs time to answer student candidate questions. In addition, the increasing of web visitors may lead to increase questionnaires and waiting times, resulting in poor client satisfaction [3].

The rapid development of computer-based information technology has made many changes in the joints of human life. One of the promising technological developments is Artificial Intelligence (AI) [4]. By using AI, the computer can perform certain tasks as performed by humans such as robot chat. Robot chat refers to as Chabot system that adopts knowledge of human nature interaction to the computer (HCI). So that computers have intuitive abilities to realize the conversation with their users using natural language [5][6]. Ranoliya el. al stated that Currently text-based Chabot is used extensively in commercial web systems [1]. Based on natural Language complexity, the implementation of Chatbot system is not an easy task because we need a set of mechanisms to extract knowledge from user FAQs [7].

In this paper describes and proposes the development of Chatbot which acts as a conversation agent that can help the role of as student candidate service. This Chatbot called DINA, Dinus Intelligent Assistance. Scope of language understood by DINA used Bahasa Indonesia because, in our university, Bahasa Indonesia is used as the official language of our daily conversation. DINA uses as a center for machine learning approach. The extracted patterns from knowledge-base then can be used to provide random responses to the user. The

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

source of knowledge based is taken from Universitas Dian Nuswantoro (UDINUS) guest book. It contains questions and answers about UDINUS, particularly related to admission questions. This knowledge based is used to store information that is frequently asked by the student candidate and search based on keywords, phrases, and examples that have been customized into database. For unanswerable question will be a new data for Chabot and will automatically be stored in the database of knowledge based. In this paper, Chatbot is needed for admission service domain in the field of education to provide student candidate satisfaction. Finally, DINA Chatbot is expected to help the admission staffs work and shorten time for student candidate to get information about UDINUS.

The rest of the paper is organized as follows: Section II presents the technical arrangement on the DINA Chatbot relies on, also provide textual entailment in Dialogflow framework. Section III introduce development the OWL ontologies to integrate the Dialogflow model. Section IV details our technical solution to extend the Dialogflow model with domain knowledge from OWL ontologies. Section V illustrates how the textual entailment is used to support conversations on DINA Chatbot.

II. RELATED WORK

The history of question answering system research and artificial intelligence produces the modern innovations in the form of chat robots derived from early systems such as ALICE [4] and ELIZA [8]. This system is built as a dialogue system or a conversation agent [9]. This conversation can use imperative or control commands such as voice-activated phone calls, music selection, or television control.

This paper [9] describe an approach to develop Chatbot for college management system using AI. The Chatbot simulates conversations in the form of computer programs that can perform dialog with users in natural language. A Chatbot is just a program, and not a robot (it does not have a body and does not have a mouth so unable to talk like a human). It is a dialog between people as users with Chatbot which is carried out by typing what will be asked and then Chatbot will respond it. The person who creates and develops a Chatbot program is called a botmaster.

A University Chatbot related to Frequently Ask Question (FAQ) was developed by [10], which is used as a question answering application about FAQ at Manipal University. The FAQ chat is information exchanges application developed using the Artificial Intelligence Markup Language (AIML) pattern that retrieves set of matching information from the FAQ databases. The Chatbot FAQs of Manipal University will assist students to get information needed like availability of services, updates regarding activities happening inside the campus, ranking of the university, university environment, and many more about academic information.

Debasatwa et.al, [11] stated in their paper that there are many types of cloud-based Chatbot platform which are api.ai now called Dialogflow, Luis.ai, Pandorabots.com, and wit.ai. In the paper, the authors compare 4 different Chatbot platforms based on Natural language processing capability and complex feature development abilities. Based on the result of Dialogflow evaluation, they selected to develop Chatbot because Dialogflow has the ability to handle sub-intent purposes of user's text as input. The intelligent Chatbot can be accessed using web browser and Android apps. The node.js runtime environment has been used for server side processing and client-side requests. The user interface has been developed used The BotUI Javascript framework.

This study [12] proposes an Ontology-based Chatbot on the E-commerce domain. The data used in the form of E-commerce Website and PROTEGE platform are used to build ontology templates that retrieve data from sources (using Jape rules). After the process of the data retrieval into the template, the user can ask questions to the bot and present the answer in text form corresponding to user input that will be set using Natural Language Processing (NLP) in Dialogflow Platform. Dialogflow uses intent and other properties built through *python.* API calls the ontology template and the query to be answered by Chatbot from data taken in the Ontology template. Based on the results, the proposed system successfully maps the relationship between multiple intents and retrieves the data according to the user's request. Base on the related work above, the author proposes to construct DINA Chatbot using Dialogflow platform and ontology knowledge.

A. Knowledge based (KB)

There is some work on the extraction of knowledge from the online web community to support Q&A and summarization Development of KB depends on the use of existing tools and information resources (eg FAQs, APIs, system logs, etc.) [13][14]. These research approach is to develop characteristic of their corpora and matched for their specific tasks, but they constrain each corpora and their tasks. Hence, it is difficult to apply their method directly to our Chatbot approach of knowledge extraction. Furthermore, DINA uses the knowledge sources, i.e., guest book from admission office of university. Data extract from guest book become a FAQs corpus. Those knowledge based corpus contains knowledge for answering student candidate problem solving. This study uses knowledge based approach and Rule-Based Reasoning which contains:

1. Knowledge in bot is represented in the form of IF-THEN or in the form of Production Rules.
2. The determination of the initial rules (rule antecedents) is carried out using inference logic.
3. The appropriate rules are placed on the agenda and can be activated so that it will generate new facts on the right side. Furthermore, activation of one rule is part of another activation rule.

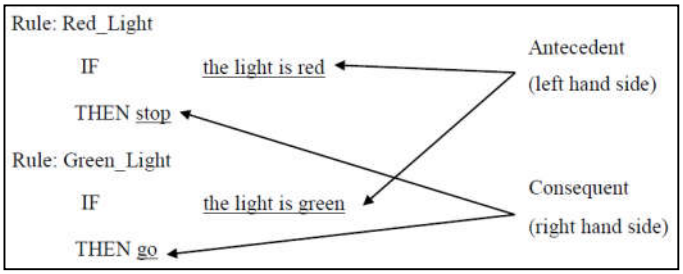


Figure 1. Knowledge based approach - Rule-Based Reasoning

B. *Dialogflow Framework*

Dialogflow is one of the Chatbots development platform based on natural language conversation [2]. When a user sends a message to Chatbot through some intermediate media such as telegram, line, and web, this message will be proccessed in the Dialogflow platform where the question will be changed in the object form. After that, it is forwarded to the external web server and undergoes some appropriate logic entailment process to complete the user request and responds to the user in text form.

The Dialogflow provides important elements to allow a conversation with the user such as:

- Dialogflow provides the dialog manager that will service us to obtain the intents of the query posted by the user using NLP.
- Create Entities to declare the various type of question (@date, @beasiswa, @nomor and etc), synonyms of the text, and properties of the various user input and output.
- Generating the JSON form node.js to recognize the user's question from Intent and entity platform
- JSON code is used to get data from ontology template PROTÉGE and sent to messaging bots which forms the interface for the project.
- Machine learning approach is used in the program that acts as an advisor or smart consultant by taking knowledge stored in a particular domain.

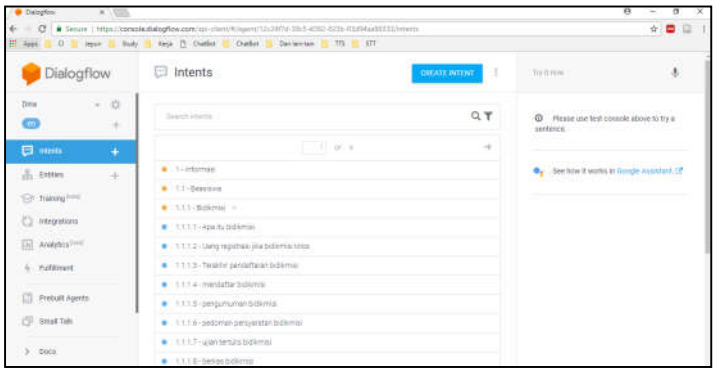


Figure 2. Developing Intents and entities for DINA Chatbot

Figure 2 shows the functionality in Dialogflow. It presents when a user enters data in an Dialogflow platform. It is first checked if it matches a predefined purpose then search and matching the input of the user with intents. Dialogflow has a feature called "Default Fallback intent" to handle user input that does not match pre-defined intent. User speech activation can be conditioned by the use of context, resulting in more human-like conversations.

C. *Ontology template*

Several NLP research has adopted the advantage of ontology such as Chatbot [14], sentiment analysis [15], and question answering[15]. Moreover, our previous study related to ontology also have been published in several international journal and conferences [16][17][18]. Commonly, there are two ways of ontology development that are using manual and automatic manner. Manual ontology development will take a lot of time and effort as compared to development automatically because we have to detail in all the corpus we will use.

There are several steps in how to build an ontology: First we have to select a domain for our ontology. In this step we choose admission knowledge based on the university registration information. Knowledge based that we used is from Frequently Asking Question (FAQ). We divide into several part along with category of FAQ. Each FAQ consists of several question and answer, so to make it easy we group one question with the answer into one intent. After that, step continued to build a glossary of terms. Means each intent have its special term to make it easier finding the intent and for categorizing. Third, build concept hierarchy. As explained before, each group of FAQ divided into its own intent so there are child of knowledge from each group. To make it as hierarchy of knowledge, index must be put for each intent created. Finally after all one knowledge of registration information are describe into hierarchy of knowledge divided into each group and has its own index, so Chatbot ontology for admission of the university was created showed in figure 3.

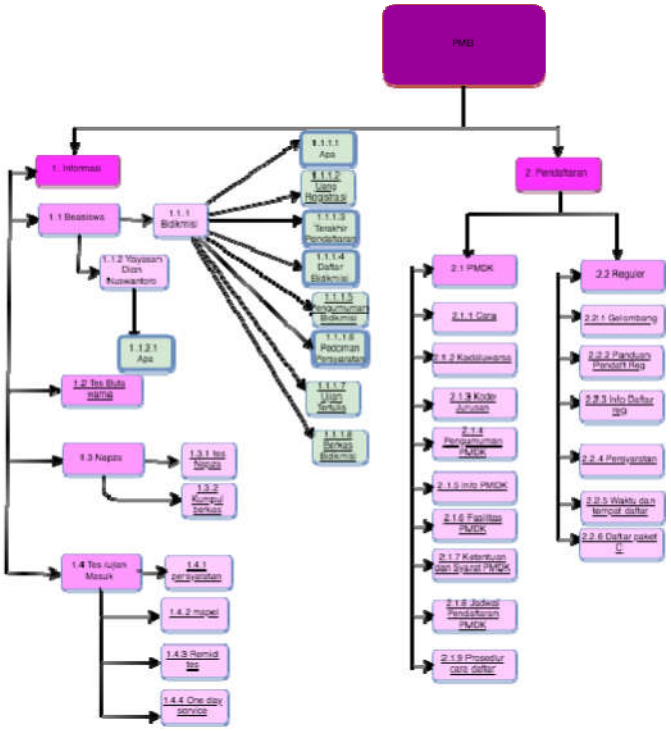


Figure 3. Admission Chatbot Ontology

III. CHATBOT ARCHITECTURE

The DINA Chatbot takes input through social media platform i.e., telegram and line, or website. It analyses that input using NLP provided by Dialogflow to find out what the user is trying to ask and responds accordingly. DINA uses a

modular architecture to respond to user input. Each module contains knowledge based, initialization mechanism, and logic to handle user requests. DINA also use a modular architecture to respond to the user input. Each module includes knowledge based, initialization mechanism and logic to handle user requests.

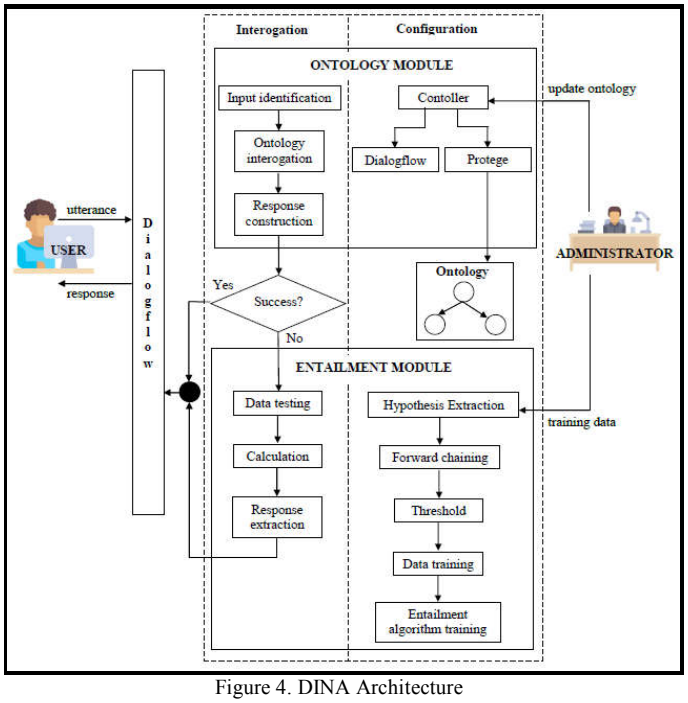


Figure 4. DINA Architecture

Figure 3 shows that DINA comes with two main modules: the learning module created using Forward chaining, which is used as the default fallback mechanism (providing answers to not pre-defined user questions), and the ontology module, which is used as an answering machine for user questions based on data stored in the OWL. DINA ontology and corpus consist of question answering argument. The modular architecture ensures that the system tries to answer questions using information in ontology first and only if no answer can be generated entailment module is used. The following sections describe the module of Chatbot in detail.

The Chatbot module will be divided into two parts. The first part is the integration part used to process user requests. This section will send a query that inputs the user to the conversation agent and continues processing phases using Dialogflow. Platform module in the Dialogflow will perform a search to customize the input text with an existing intent in the data corpus. The second part is the configuration section. Therefore, both modules can communicate and interact with the web server. The configuration part of one of the modules must include logic to handle the Dialogflow model used (in our system this is done by the ontology module).

A. Ontology module

This module is used to extract knowledge from the corpus into the ontology format. Hence, it will produce the suitable answer with *entities* and translate from otology template to Dialogflow format. The following step of making ontology as a notation language:

1) Configure Dialogflow by creating classes, properties, name of individual, and defines ontology. So it expands the knowledge based of Dialogflow
2) During the process of translation, the placeholder parameters on the ontology replaced by entities in Dialogflow
3) Create a set of simple entities for each class, and the combined entities are required if a class has subclasses
4) Create a synonymous concept for each class and add in the class entity.
5) Creation of parameters related to the entity class to facilitate responding of user input. All templates associated with labels on ontology have the same number of parameters.

B. Learning module

Forward chaining has been applied to the DINA Chatbot as learning module. The process of forwarding chaining begins when user entering information or questions into the text form provided by the application. Figure 4 show the flowchart of learning module.

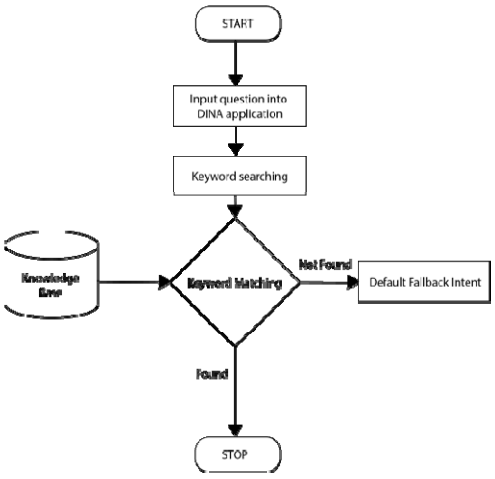


Figure 5. Flowchart Forward Chaining DINA Learning Module

Based on figure 4, the question is split into several words and searched for keywords that match the one in the application. After the keywords found, then the application answers the user's question with the correct answer. If the application can't find the keyword then the answer appears which is taken from Default Fallback intent .

Technically, the textual entailment is enabled either when Dialogflow does not identify the intent or when due to restrictions, no response templates can be used in the ontology request module. It will send default answer or specified default intent fallback, Extraction of the hypothesis based on hierarchical grouping able to estimate the similarity between user input and debate questions.

IV. RESULT

The Chatbot is implemented to meet the academic information needs of new student candidates. Chatbot was

built using Dialogflow platform for Universitas Dian Nuswantoro. It will help potential students to take information such as admission procedures, university ranking and study program accreditation, service availability, university environment and so forth. The implementation of forwarding chaining ontology based Chatbot application is located in the search process for answering questions from the user. The first process occurs when the user enters the question into the question box. After that, the input will be searched based on the selected category and the keywords in question as described in Figure 4. Here is a table of the implementation results in DINA Chatbot.

Table 1. Testing Result

No	List of question	Keyword/ Intent	Answer	Response	Correct Response
1	*Di mana kah letak kampus Udinus?* (Where is Udinus campus?)	*Lokasi Alamat Udinus*	*Kampus I: Jalan Nakula I No. 5 - 11 Semarang dan Kampus II: Jalan Imam Bonjol No. 207 Semarang.* (Campus I: Nakula I Street No. 5 - 11 Semarang and Campus II: Jalan Imam Bonjol No. 207 Semarang.)	Correct	-
2	*Bagaimana cara mendaftar di Udinus?* (How to register at Udinus?)	*Pendaftaran*	*Pendaftaran Udinus ada 5 macam, antara lain: PMDK, Reguler, Online, Transfer, dan Pindahan. Silahkan ketik "Pendaftaran PMDK" atau "Syarat daftar reguler", atau "Prosedur daftar transfer", dan lain sebagainya. Kamu ingin lewat jalur apa untuk daftar udinus?* (Registration Udinus there are 5 kinds, among others: PMDK, Regular, Online, Transfer, and Moving. Please type "PMDK Registration" or "Regular list requirement", or "Transfer list procedure", and so forth. What do you want to go through on the registration udinus?)	Correct	
3	*Berapa biaya pendaftaran di Udinus?* (How much is the registration fee at Udinus?)	*Uang pendaftaran*	*Uang pendaftaran Udinus sebsar Rp.200.000* (Udinus registration fee is Rp.200,000)	Correct	-
4	*Biaya perkuliahan di Udinus?* (Tuition fee at Udinus?)	*Biaya*	*Biaya pendidikan ada di web. Silakan dibaca di sini: http://id.dinus.ac.id/fees* (The cost of education is on the web. Please read here: http://id.dinus.ac.id/fees)	Correct	-
5	*Min, itu yg surat pernyataan bebas narkotika wajib?* (admin, is the narcotic-free letter a mandatory?)	*Tes Napza*	*Tes bebas narkoba merupakan salah satu syarat wajib registrasi di Udinus. Tes bisa dilakukan di klinik terdekat rumah Anda atau di Poliklinik Udinus (depan Gedung C) pada hari Senin-sabtu, 07.30-16.00 WIB. Registrasi saja terlebih dahulu. Setelah mendapatkan NIM, barulah ke poliklinik. Biaya tes sekitar Rp. 70.000.* (The drug-free test is one of the requirements for registration in Udinus. Tests can be done at your nearest clinic or Udinus Polyclinic (in front of Building C) on Monday-Saturday, 07.30-16.00 WIB. Register first. After getting the NIM, then to the polyclinic. The test fee is about Rp. 70,000.)	Correct	-
6	*min apa ada perbedaan dari pendaftargel IIA dengan IIB?* (min what is the difference between IIA gel registries and IIB?)	*Tidak ikut matrikulasi*	*Matrikulasi adalah fasilitas untuk maba pendaftar gelombang awal, sehingga yang tidak ikut matrikulasi tidak akan mendapatkan kegiatan pengganti. Kegiatan selanjutnya yang wajib diikuti semua maba adalah dinus inside. Untuk melihat detail infomasi pelaksanaan dinus inside silahkan akses http://dinus.ac.id/* (Matriculation is a facility for new student from first batch, for who do not participate in matriculation will not get replacement activities. The next activity that must be followed by all new student is dinus inside. To see detail information on the implementation of the dinus inside please access http://dinus.ac.id/)	Not corrrect	perbedaan hanya pada tanggal pendaftaran. (The difference only on registration date.)
7	*Kak di udinus ada kelas karyawan?* (Sister in udinus there is employee class?)	*Kelas karyawan*	*Kelas karyawan hanya ada di Fakultas Kesehatan, program studi Kesehatan Masyarakat S1 dan Rekam Medis D3. Kuliah di hari jum'at mulai pukul 14.00 WIB dan di hari sabtu mulai pukul 08.00 WIB.* (Employee class only available in Faculty of Health, Public Health Study program S1 and D3 Medical Record. Lecture on Friday starting at 14.00 pm and on Saturday starting at 08.00 pm.)	Correct	-
8	*Mau tanya, kalo saya dari IPA tp ngambil jurusan Ilkom apa boleh?* (Want to ask, if I'm from IPA but wanna take majors Ilkom, is it possible?)	*Pindah jurusan*	*Jika Anda belum melakukan tes, ganti jurusan bisa dilakukan di petugas pendaftaran. Jika sudah melakukan tes, ganti jurusan dilakukan di bagian registrasi. Jika sudah menjadi mahasiswa aktif di Udinus silakan konsultasi ke Kepala Progdi terlebih dahulu.* (If you have not done the test, change majors can be done at the front officer. If you have done the test, change majors done in the registration. If you are already an active student in Udinus please consult the Head of Faculty first.)	Not correct	Boleh. (Yes, you can)
9	*Min,saya mau nanya. Buat pendaftaran tahun depan dri kejar paket apa udinus mau menerimanya?* (Min, I want to ask. I want register next year through the equivalence test certificate,	*Kejar paket*	*Bisa, asalkan sudah ada ijazah. Silakan disimak info berikut untuk jadwal dan persyaratan: http://admisi.dinus.ac.id/pmbonline/menu/jalurpendaftaran.* (Yes, as long as you have a certificate of graduation. Please check the following info for schedule and requirements: http://admisi.dinus.ac.id/pmbonline/menu/jalurregistrasi)	Correct	-

	is Udinus willing to accept it?)				
10	*Min, persyaratan lengkap bidikmisi apa saja? Trimakasih* (Min, what are the complete requirements of bidikmisi? Thank you)	*pedoman persyaratan bidikmisi*	*Pedoman/ persyaratan lengkap dapat diunduh pada website ini bidikmisi.belmawa.ristekdikti.go.id/petunjuk/pedoman* (Full guidelines / requirements can be downloaded on this website bidikmisi.belmawa.ristekdikti.go.id/instructions/guide)	Correct	-

The testing process is done in accordance with the form of questions that are often asked by the user. This stage is done to determine the ability of the application in handling questions. The testing process was more focused on the sentence form which used 166 intents. It is done by running the application DINA then enter the sentence in the text field that has been available.

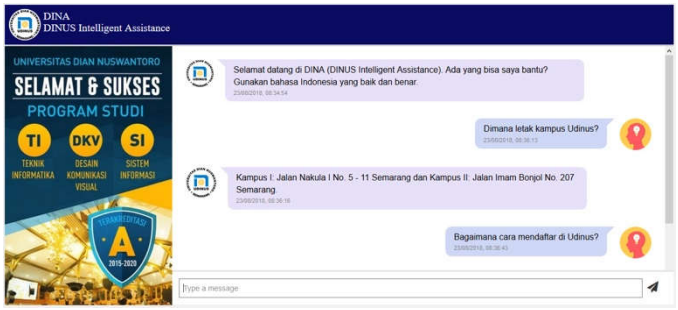


Figure 6. DINA Display in website

From 166 intents, The author tested ten random sample questions. The result showed that it got eight questions answered correctly and able to answer any question from perspective student as showed in Figure 5.

V. Conclusion

Through DINA Chatbot that researcher proposes, the conclusion drawn is that the new proposed system will help users to obtain information needed. With the proposed approach, major flaws existing in current Chatbots that cannot understand the relationship between entities and properties can be solved. The proposed system successfully maps and retrieves data. By using this study, we can further build an intelligent chat bots to help users find the appropriate information. It also help user to obtain answer without waiting for the admin staff's answer. With enough training from ontology bots, we can now successfully interact with ontology bots. For further development, we will equip our Chatbot with voice recognition and text-to-speech capability.

References

[1] B. R. Ranoliya, N. Raghuwanshi, and S. Singh, "Chatbot for university related FAQs," *2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017*, vol. 2017–Janua, no. September 2017, pp. 1525–1530, 2017.

[2] D. Toniuc and A. Groza, "Climebot: An argumentative agent for climate change," *Proc. - 2017 IEEE 13th Int. Conf. Intell. Comput. Commun. Process. ICCP 2017*, no. October, pp. 63–70, 2017.

[3] Z. N. Baiti and F. Nugroho, "Aplikasi Chatbot 'MI3' Untuk Informasi Jurusan Teknik Informatika Berbasis Sistem Pakar Menggunakan Metode Forward Chaining," *Matics*, vol. 5, no. 3, p. 9, 2013.

[4] J. Hill, W. Randolph Ford, and I. G. Farreras, "Real conversations with artificial intelligence: A comparison between human-human online conversations and human-Chatbot conversations," *Comput. Human Behav.*, vol. 49, pp. 245–250, 2015.

[5] B. Hettige and A. S. Karunananda, "Octopus A Multi Agent Chatbot," *Proc. 8th International Research Conference, KDU, Published November 2015*, no. November, pp. 41–47, 2015.

[6] A. Argal, S. Gupta, A. Modi, P. Pandey, S. Shim, and C. Choo, "Intelligent travel Chatbot for predictive recommendation in echo platform," *2018 IEEE 8th Annu. Comput. Commun. Work. Conf.*, pp. 176–183, 2018.

[7] A. Sangroya, P. Saini, and C. Anantaram, "Chatbot as an Intermediary between a Customer and the Customer Care Ecosystem," *Proc. 9th Int. Conf. Manag. Digit. Ecosyst. - MEDES '17*, pp. 128–133, 2017.

[8] H. Shah, K. Warwick, J. Vallverdú, and D. Wu, "Can machines talk? Comparison of Eliza with modern dialogue systems," *Comput. Human Behav.*, vol. 58, pp. 278–295, 2016.

[9] P. K. Bala, M. Kumar, S. Hulawale, and S. Pandita, "Chat-Bot For College Management System Using A . I," 2017.

[10] B. R. Ranoliya, N. Raghuwanshi, and S. Singh, "Chatbot for university related FAQs," *2017 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2017*, vol. 2017–Janua, no. January, pp. 1525–1530, 2017.

[11] D. Dutta, "Developing an Intelligent Chat-bot Tool to assist high school students for learning general knowledge subjects," p. 13, 2017.

[12] A. Vegesna, "Ontology based Chatbot (For E-commerce Website)," vol. 179, no. 14, pp. 51–55, 2018.

[13] R. Kincaid and G. Pollock, "Nicky: Toward a Virtual Assistant for Test and Measurement Instrument Recommendations," *Proc. - IEEE 11th Int. Conf. Semant. Comput. ICSC 2017*, pp. 196–203, 2017.

[14] A. Hallili, "Toward an Ontology-Based Chatbot Endowed with Natural Language Processing and Generation Amine Hallili To cite this version : HAL Id : hal-01089102," p. 7271, 2014.

[15] K. K. N. -L S Chetan Rao, Dinesh Kini, Kanthraj S,

"Chatbot-a Java Based Intelligent Conversational Agent," *Int. Res. J. Eng. Technol.*, vol. 4, no. 4, pp. 3575–3578, 2017.

[16] H. A. Santoso, J. Zeniarja, A. Luthfiarta, and B. J. Wijaya, "An ontological crawling approach for improving information aggregation over eGovernment websites," *J. Comput. Sci.*, vol. 12, no. 9, pp. 455–463, 2016.

[17] H. A. Santoso, S. C. Haw, and Z. T. Abdul-Mehdi, "Ontology extraction from relational database: Concept hierarchy as background knowledge," *Knowledge-Based Syst.*, vol. 24, no. 3, pp. 457–464, 2011.

[18] H.-A. Santoso, S.-C. Haw, and C.-S. Lee, "Software Reuse: MDA-Based Ontology Development to Support Data Access over Legacy Applications," *Softw. Eng. Comput. Syst.*, pp. 130–142, 2011.

Comparison of Severity on Mobile Government Application Mobile

Rully Alexander
Magister Teknik Informatika
Universitas Atma Jaya Yogyakarta
Yogyakarta, Indonesia
rully.alexander18@gmail.com

Andhika Galuh Prabawati
Magister Teknik Informatika
Universitas Atma Jaya Yogyakarta
Yogyakarta, Indonesia
andhika.prabawati@gmail.com

Djoko Budiyanto Setyohadi
Teknik Informatika
Universitas Atma Jaya Yogyakarta
Yogyakarta, Indonesia
djoko.bdy@gmail.com

Abstract- At present E-Government in every country has been growing and experiencing stable growth, it is seen from many E-Governments which have started to provide services and information through an application especially with a mobile application. Through this development, we can get the benefit from user engagement in using a good E-Government app, usability, and credibility believed to be one of the factors affecting the developmental and growing participation.

This research presents an empirical study in evaluating the usability of mobile E-Government mobile applications between two countries, namely BMKG INFO and MET Service Singapore. This study was conducted to determine the severity of the two applications. The results obtained show that by knowing the severity of the application, can be used for improvement on them.

Keywords-Severity, Heuristics, Mobile Application, E-Government.

I. INTRODUCTION

Mobile devices are the favorite gadget in the broader community. In recent years, mobile devices have advanced rapidly in use by many users. Mobile phones have become the most popular and widespread personal user device. This condition poses a new challenge to its usefulness [1]. The use of cellular connections around the world has increased by nearly 3.3 billion per day [2]. Currently, many online applications add features which are implemented on mobile devices, and it is updated according to user needs usually. According to this situation, some testing is needed particularly for usability testing.

Usability testing generally consists of several evaluations of the software based on the proper considerations of the functionality. However, usability testing typically scares computers and mobile applications that involve the usage of the applications in the development process [3]. The results of usability evaluation can be crucial to the success of mobile applications, high usability levels suggest reducing development costs, lowered maintenance costs, and higher customer retention [4]. In additional the characteristic, an interface can influence the perception of a user. Therefore usability it's an implication in mobile design [5] Many techniques are used to measure usabilities, such as measuring user satisfaction through post-test questionnaires such as SUS [6] and SUMI [7].

Heuristic evaluation is a favorite way to evaluate usability [8], the advantages of heuristic evaluation methods are fast, cheap and easy to find problems with usability [1]. However naturally, evaluation cannot be performed automatically and is often very broad and subject to the evaluator's understanding [9]. In a heuristic evaluation, evaluators evaluate the usability of adherence to interface design over a set of heuristic uses to identify potential usability issues and to measure usability [10].

In this research will conduct usability evaluation on mobile application of E-Government weather of different country by using a heuristic method. It is expected that the results of this research can know the severity usability of the interface design in mobile E-Government mobile application between different countries. Also, the evaluation will be done to find out how big the problem users find in the app. In the future, the application can perform improved look and functionality, especially for student level users.

In this paper, a comparison of two m-government applications is presented, especially in the fields of meteorology, climatology, air quality and geophysics. This comparison is useful since the comparison can be used by the community to improve its application and then the disaster response can be refined.

II. BACKGROUND STUDY

A. E-Government Application

Currently, E-Government is highly developed E-Government can be interpreted as internet users, for example in web technology that is a tool in delivering information and providing government services to its users [11]. E-Government is a form of use in today's information and communication technology to enable governments to become more efficient, cost-effective, and participative, to facilitate government services to be more convenient, allowing the public to access information more and making governments more accountable to citizens country [12]. At present every E-Government has implemented and built their E-Government system, although it is rapidly evolving and yet some studies address issues. For example, studies using existing E-Government comparisons exist in Indonesia, the results show that local government websites are still in the making, refinement, and interactive phase [13].

B. Usability

Testing or usability measurement is a series of activities undertaken to measure the evaluation of efficiency, effectiveness, and satisfaction of software users. Two events can be done in evaluating usability, the first is to use the opinions of the software developers and the second is to use usability metrics [14]. Usability evaluation can usually be defined as a procedure developed by well-defined activities to collect data related to end-user interactions and software products, to determine how specific parts of the software contribute to achieving a goal [15]. Usability evaluation is also used for the development of a software [16].

C. Heuristic Evaluation

Heuristic evaluation is a way of finding out the problems that exist in software or on display. The question then relates to the difficulty level of the problem. In the heuristic method, there are ten primary principles in the application of heuristics that can be used as a guide in performing or measuring usability. These ten principles or guidelines are described in a previous study which discusses how the concept of the heuristic method is used in the process of evaluated interface usability [17].

Currently, the heuristic can be used to evaluate an interface a heuristic adjustment is required to ensure usability issues specific to user interactions and not negligible. From the 80's, several authors have developed and published a new heuristic section with guidelines for design, usability, ergonomics, software engineering, and among other fields related to human and computer interactions [18].

D. Mobile Application

Mobile apps are a set of programs that run on mobile devices and perform specific tasks for users. Mobile apps are a new and very rapid part of developing global information and communication technologies. Mobile apps have broad usability for common areas such as calling, messaging, browsing, chatting, social networking communications, audio, video, games, etc. [19]. Today, mobile applications and services are increasingly helpful in everyday life situations, meeting our needs for information, communication, entertainment or recreation [20].

III. METHODOLOGY

In this research, heuristic Neilsen method has been applied to fulfill the purpose of the study. The initial phase of the study begins by making questionnaire questions according to the used-process. In this case, the questionnaire was adapted to the fundamental principle in Nielsen heuristic methods then it is adapted to mobile application usability areas [20].

A. E-Government Weather Application Selection

Two mobile E-Government mobile applications from different countries are selected to target this research (1). Info BMKG (Meteorology, Climatology, and Geophysics), Indonesia is an application that displays weather, climate, air quality and earthquake estimates within three days in Indonesia. 3-day weather forecast infograph in the morning, noon, night and morning in all districts and cities in Indonesia. In the BMKG application menu, is displayed four main information, in the weather forecast, BMKG info, whether the early warning, and current earthquake information and (2). The second app is Weather @ Sg from Singapore. This application is often called The Meteorological Service Singapore (MSS). MSS is Singapore's national authority to deal with weather and climate. MSS is the operating division under the national environmental agency (NEA). Its mission is to observe and understand the weather and climate affecting Singapore and provide services to support national and international needs. The primary function of MSS is to collect and maintain long-term weather records at the national level, provide reliable weather and climate service information, conduct high-quality research to advance climate and climate prediction, and assess risk and harmful impacts the natural environment.

B. Usability Questionnaire

The questionnaire is used to ensure the questions given to the respondents are the same, and the responses obtained are easily obtained quickly. Each respondent is asked to fill in each item, and each question has five points of answers assessed using a Likert scale. The Likert scale, we can distinguish arguments on each item [21]. The Likert scale uses an odd number of responses that can enable participants to provide opinions with positive, neutral or negative choices. After that, the information obtained by using the Likert-scale five points can be easily collected and analyzed (sample question from the questionnaire can be seen from Figure 1.

Assessment Score

Strongly Disagree : 1

Disagree : 2

Don't Know : 3

Agree : 4

Strongly Agree : 5

NO	HEURISTIC EVALUATION
A	**Visibility of System Status:**
A.1	Is the information already in accordance with the name of The Info BMKG application? ☐ Strongly Disagree ☐ Disagree ☐ Don't Know ☐ Agree ☐ Strongly Agree
A.2	Do you know the flow of searching for information on the Info BMKG application? ☐ Strongly Disagree ☐ Disagree ☐ Don't Know ☐ Agree ☐ Strongly Agree

Fig 1. Sample Questionnaire

C. Participants

The distribution of questionnaires was conducted in Yogyakarta, Indonesia. Forthy (40) participants are students from 3 universities in Yogyakarta. This study is an exploratory study explaining the number of participants who somewhat in this study conducted participants voluntarily limited. In this research, the participants come from 3 universities in Yogyakarta, Indonesia is Atma Jaya University Yogyakarta, Gadjah Mada University Yogyakarta, and Duta Wacana Christian University. Before the participants fill out the existing questionnaires, the researcher gives an explanation and a general description of the research objectives.

D. Evaluation Procedure

In this study, 40 participants were given the task to try to interact with two mobile applications that became the purpose of this study. Each evaluation is done in several stages, in evaluating the interaction using tasks and questionnaires. Reviewing the app as a whole lets participants view and use mobile apps or focus on design elements on specific interfaces in the app. With this the participants who are new to using this application for the first time get initial interaction and shape their common perception can be formed.

After that, the participants were asked to interact by following the guidelines already in the questionnaire on the mobile application that was the target of the research. To that end, the task sheets adjusted to the benchmark 10 Nielsen heuristic principles are given to the participants. The participants were asked to perform tasks according to the questionnaire one by one indefinitely. After the task is completed, participants are asked to fill out a Nielsen usability heuristic evaluation questionnaire.

E. Data Analysis

To conduct a user perception analysis of the evaluation of usefulness of mobile E-Government mobile application being targeted, questionnaires and observations data by using Likert scale with "strongly disagree" one point, "disagree" two points, "do not know" three points, "agree" four points and five points for the statement "strongly agree". Then calculate each score from the average of all heuristic elements for each evaluation performed. The severity rating is given based on the average result. Table 2 shows the severity rating classification which has been obtained in this study.

TABLE I. CLASSIFICATION OF SEVERITY RATING

Severity rating	
0% - 19.99%	4-catastrophe
20% - 39.99%	3-major problems
40% - 59.99%	2-minor problems
60% - 79.99%	1-cosmetic problems
80% - 100%	0-perfect

After the data were collected based on the questionnaire, the average score for both apps evaluated for use was noted. The average value is grouped into ten heuristic principles. In table 2 is the average value already processed in the study.

TABLE II. MEAN APPLICATION EVALUATION SCORE

Heuristic	Mobile Application Weather E-Government	
	Info BMKG, Indonesia (%)	***weather@sg, Singapura (%)***
H1	82.25	79.5
H2	78	73.5
H3	64.3	61.7
H4	79.5	76
H5	66.5	60.5
H6	77.17	71.50
H7	66.25	53.08
H8	57.8	72.8
H9	66.17	55.83
H10	51.5	43
Average	68.94	64.78
Rating	Cosmetic problems	Cosmetic problems

The next stage is to conduct an analysis of results based on data processing that has been done and obtained. The details will be explained in the next section

IV. RESULTS AND DISCUSSIONS

In this section presents and discusses the results of research that has been done. It includes a descriptive analysis of information that has been obtained from research on participants in a usability test assessment using the heuristic method of two mobile E-Government mobile applications.

A. Descriptive Information

Data from Table 3 shows the results of the analysis conducted based on the questionnaire data that has been filled by the participants. In processing our data using the Likert scale [22], after which the existing severity rating is obtained based on average results [24]. Furthermore, the severity rating classification analysis is determined according to what is in table 2. Where the average score of 0% - 19.99% means having catastrophe severity, 20% - 39.99% have the severity of major problems, 40% - 59.99% have severity minor problems, 0% - 79.99% have severity cosmetic problems, and 80% - 100% have perfect severity.

B. Usability Review

Based on the results of the analysis, mobile application weather E-Government info BMKG obtained the highest average value. Info BMKG has an average value of 68.94%, which is the severity of cosmetic problems, while the weather @ sg has an average value of 64.78% which is also the severity of the severity of cosmetic problems. Based on the severity obtained and presented by the researcher, for the severity of cosmetic problems, there is no need for significant improvement unless it has a chance to improve it. In this study, heuristic evaluation is conducted on two different mobile Internet E-Government application interfaces, so there

is an opportunity to make and make comparisons between the two applications. We can get an improvement area based on that comparison. Figure 2 shows a graph depicting the contrast of the mean scores for both mobile weather apps.

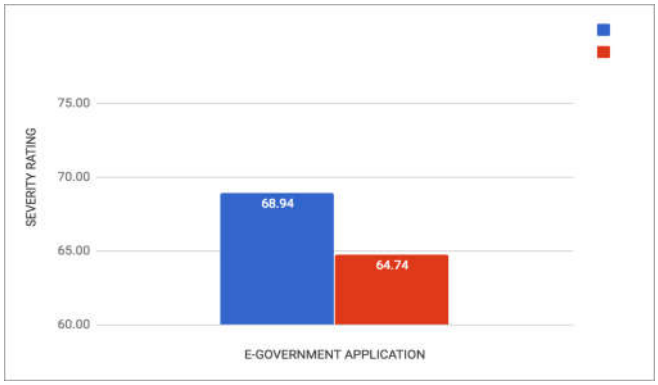


Fig. 2. Evaluation Result

1) Mobile Application Weather - INFO BMKG (Indonesia)

Based on the average data obtained after the processing of questionnaire data with the Likert scale which then calculated the average score on each element of heuristic evaluation then mobile application info BMKG info obtained a total average score of 68.94% which is the severity of the severity "Cosmetic Problems." Score value on heuristic element tested on BMKG mobile application has the highest average heuristic score 1 (H1: 82.25%) which is the score of the first element of heuristic that is visibility of system status, average score reading the lowest heuristic element in the mobile application of the weather info BMKG is in heuristic 10 (H10: 51.5%) which is the value score of the heuristic element "Help and Documentation". BMKG's mobile weather info app has the lowest comparable value score only on heuristic element 8 (H8: 57.8%) in comparison with the mi app obtained by questionnaire with Likert scale. In mobile apps weather @ sg (H8: 72.8%) which is the value score of element "Aesthetic and minimalist design". In this case, if BMKG can be better in terms of aesthetic and minimalist design it is especially useful for people in Indonesia, especially in looking for weather and disaster information on BMKG applications

2) Mobile Application Weather - @Sg (Singapore)

In this application, we also perform average score calculations on each heuristic evaluation element following the data we obtain through a questionnaire with a Likert scale. In mobile apps, the weather @ sg earned an average total score of 64.78% which is the severity of the "cosmetic problem" severity. Score values on heuristic elements tested in mobile apps weather @ sg have the highest average heuristic score 1 (H1: 79.5%) which is the value score of the heuristic element "visibility of system status", this application has an average reading the lowest score on the heuristic element of the mobile application of the weather @ sg is in heuristic 10 (H10: 43%) which is the heuristic value score of "Help and Documentation". mobile application weather @ sg has the highest comparative value score only on heuristic element 8 (H8: 72.8%) in comparison with BMKG mobile weather info application (H8: 57.8%) which is the value score of element "Aesthetic and minimalist design."

C. Managerial Implication

Other studies that do usability testing with heuristics on M-Government applications show that the heuristic method can enable a more precise and more in-depth examination [25]. Using the principle of the heuristic method can be known some usability probems. In this study, we investigate the problem which present the usability part of the in the application. Its problems are seen in a usability level which has a low score level.

Then, after we analyze the data that has been collected then obtained a score level for each heuristic element in the application usability section tested. The results of the data analysis show that both mobile E-Government applications are examined, based on existing usability test guidelines, it is known which parts have positive and negative scores on applications such as previous research [26]. With the result data analysis shows how the results of usability testing on both applications so that can be obtained comparison usability test against the mobile application of government between two countries in contrast by using necessary severity rating.

V. Conclusion

In this study there are two subjects examined; Mobile apps weather Info BMKG, Indonesia and Mobile weather apps weather @ sg, Singapore. The results show that M-Government applications in the test have the same severity level that has the severity level of the severity of "cosmetic problem." However, the mobile application of the weather BMKG Info, Indonesia has a better total score compared to the total score of 68.94%, while the mobile weather application @ sg, Singapore has a lower score with a total score of 64.78%. In the other side, Indonesia show that it has the highest average value reading on the first heuristic principle of visibility of systems status and the lowest on the tenth heuristic principle of help and documentation. While the mobile weather apps @ sg, Singapore show he has the highest average reading on heuristic 1 (visibility of systems status) and also the lowest on heuristic 10 (help and documentation). From the results obtained it can be seen areas of improvement that can be done in the future in these two applications.

References

[1] Wangenheim CG Von, Witt AT, Borgatto AF, et al. An Usability Score for Mobile Phone Applications based on Heuristics. 2016;8(1):23-58. doi:10.4018/IJMHCI.2016010102.

[2] A. Bin Hussain, S. Aldeen, A. Abbas, M. S. Abdulwaheed, R. G. Mohammed, and A. Abdullah Abdulhussein, "Usability Evaluation of Mobile Game Applications: A Systematic Review," *Int. J. Comput. Inf. Technol.*, vol. 4, no. 3, pp. 2279–764, 2015.

[3] B. Zaid, R. Jamaluddin, and B. Wafaa, "A Comparative Study of Usability Methods for Mobile Applications," *Int. J. Sci. Eng. Res.*, vol. 3, no. 8, pp. 1–4, 2012.K. Elissa, "Title of paper if known," unpublished.

[4] B. Igler, T. Braumann, and S. Böhm, "Evaluating the Usability of Mobile Applications," *Int. J. Bus. Manag. Stud.*, vol. 5, no. 1, pp. 92–102, 2013.

[5] Setyohadi D.B., Kusrohmaniah S., Christian E., Dewi L.T., Sukci B.P. (2017) M-Learning Interface Design Based on Emotional Aspect Analysis. In: Basu A., Das S., Horain P., Bhattacharya S. (eds) Intelligent Human-Computer Interaction. IHCI 2016. Lecture Notes in Computer Science, vol 10127. Springer, Cham

[6] Brooke, J., 1996, SUS-A Quick dan Dirty Usability Scale, Usability Evaluation in Industry, vol. 189, pp. 4-7.

[7] Kirakowski, J., and Corbett, M.. 1993. SUMI: the Software Usability Measurement Inventory. British Journal of Educational Technology, 24, (3), pp. 210-212.

[8] Nielsen, J., 1994, Usability Inspection Method, John Willey & Sons, New York.

[9] K. Thitichaimongkhol and T. Senivongse, "Enhancing usability heuristics for Android applications on mobile devices," *Proc. World Congr. Eng. Comput. Sci.*, vol. I, 2016.

[10] Nielsen, J., 1993, Usability Engineering, Academic Press, New Jersey

[11] A. Muir and C. Oppenheim, "developments worldwide I : electronic government," vol. 28, no. 3, pp. 173–186, 2002.

[12] J. M. Heffron, "E-Government and the applications of technology to government services," *Basin Res.*

[13] N. Nurdin, R. Stockdale, and H. Scheepers, "Benchmarking Indonesian local E-Government," *16th Pacific Asia Conf. Inf. Syst. (PACIS 2012)*, pp. 1–13, 2012.

[14] Tullis, Tom, and Albert, Bill. 2008. Measuring The User Experience. Morgan Kauffman.

[15] A. Fernandez, E. Insfran, and S. Abrahao, "Usability evaluation methods for the web: A systematic mapping study," *Inf. Softw. Technol.*, vol. 53, pp. 789–817, 2011.

[16] F. Paz and J. A. Pow-Sang, "A systematic mapping review of usability evaluation methods for software development process," *Int. J. Softw. Eng. its Appl.*, vol. 10, no. 1, pp. 165–178, 2016.

[17] R. da S. de Q. Pierre, "Heuristics in Design: A Literature Review," *Procedia Manuf.*, vol. 3, no. Ahfe, pp. 6571–6578, 2015. (Available at : https://www.nngroup.com/articles/ten-usability-heuristics/).

[18] R. Islam, R. Islam, and T. A. Mazumder, "Mobile Application and Its Global Impact.," *Int. J. Eng. Technol.*, vol. 10, no. 6, pp. 104–111, 2010.

[19] M. G. Zaragoza, H. Kim, and D. S. Han, "Mobile Application Development , Approaches , Advancement and Process," vol. 10, no. 10, pp. 79–88, 2016.

[20] K. Wac, S. Ickin, J.-H. Hong, L. Janowski, M. Fiedler, and A. K. Dey, "Studying the Experience of Mobile Applications Used in Different Contexts of Daily Life Categories and Subject Descriptors," *Proc. first ACM SIGCOMM Work. Meas. up stack*, pp. 7–12, 2011.

[21] Fung, R. H. Y.*, Chiu, D. K. W., Ko, E. H. T.*, Ho, K. K., & Lo, P. (2016). Heuristic Usability Evaluation of University of Hong Kong Libraries' Mobile Website. The Journal of Academic Librarianship, in press.

[22] G. Norman, "Likert scales, levels of measurement and the 'laws' of statistics," *Adv. Heal. Sci. Educ.*, vol. 15, no. 5, pp. 625–632, 2010.

[23] Severity Rating for Usability Problems: http://www.useit.com/papers/heuristic/severityrating.html [Acessed May 15, 2018].

[24] Othman, Mohd Kamal, Fadhullah Mahudin, Cassandra Henry Ahaguk, and Muhd Farhan Abdul Rahman. 2015. "Mobile Guide Technologies (Smartphone Apps): Collaborative Heuristic Evaluation (CHE) with Expert and Novice Users." *Proceedings - 2014 3rd International Conference on User Science and Engineering: Experience. Engineer. Engage, i-USEr 2014*, 232–36. https://doi.org/10.1109/IUSER.2014.7002708.

[25] Garcia, A.C. B., Maciel, C., & Pinto, F. B. (2005). A quality inspection method to evaluate E-Government sites. Lecture Notes in Computer Science, 3591, 198–209

[26] Z. Huang and M. Benyoucef, "Usability and credibility of E-Government websites," Gov. Inf. Q., vol. 31, no. 4, pp. 584–595, 2014.

Prediction of Glucose Content and Level of Tomato Maturity With Backpropagation Based On Image Vector Feature

Kelvin Herwanda Tandrio, Sarngadi Palgunadi Yohanes, Esti Suryani
Informatics Department
Universitas Sebelas Maret
Surakarta, Indonesia
kelvin_ht@student.uns.ac.id, palgunadi@staff.uns.ac.id, estisuryani@staff.uns.ac.id

***Abstract*— Predicting of tomato maturity is a major factor in tomato quality. It also affects the glucose content which is a side factor. The higher the maturity level is the higher the glucose content. Predicting of tomato maturity is still done based on direct visual vision. However, the disadvantage is the subjective nature of the sorting operator that is inconsistent in the classification process. With image processing approach, it can be predicted glucose content and level of maturity tomato. This study aims to predict glucose content and maturity level of tomato with backpropagation based on image vector feature of tomato. The sample image used is the image of tomato skin. Through it is determined by using red histogram, RGB, and HSV. Determination of tomato maturity level uses backpropagation with red histogram parameter. Meanwhile, the prediction of glucose content is done by using linear regression approach with RGB and HSV color models. The results showed that the determination of tomato maturity level with backpropagation obtained an accuracy of 86.67% with 4 hidden layer neurons and MSE = 0.01. The determination of glucose content has an error value of 7.57% calculated from the regression equation that is $Y = 3.82505 + 0.00598(|R - G|) + 0.715703\left(\frac{|R-G|}{R+B}\right)$.**

Keywords*— tomato, glucose, maturity, image processing, RGB, HSV, Backpropagation

I. INTRODUCTION

Tomato (*Lycopersicon esculentum*) is one of the most popular, versatile, and many consumed vegetables in the world [1]. Tomatoes have a lot of nutrients. There are very useful for body health such as the content of beta-carotene as a form of vitamin A for eye health, vitamin C content to prevent thrush and others [2]. Tomato is a favorite type of vegetables many consumed by the people of Indonesia. This tomato is the favorite vegetable category of the five most favorite vegetables in 2015 - 2016 [3].

Tomato maturity is a significant factor in determining the quality of tomatoes. Identification and detection of tomato maturity is still done based on direct visual vision. However, the disadvantage is the subjective nature of the sorting operator that is inconsistent in the classification process [4]. Therefore, it takes technology and methods to facilitate the process of classification and consistency. With the development of current technology, it makes it easy for people, including farmers, to identify tomato maturity. [5]. Tomato maturity level also affects the content of glucose which is a side factor to know the quality of tomatoes. The higher the level of maturity, the higher the sugar content [6].

Determination of tomato maturity level has been developed by several researchers. One of them is the determination of maturity level and disease in tomato using Discriminate Clustering Method [7]. This research made three classifications that are raw, mature, and too mature [7]. The analysis process uses a Gaussian filter to remove noise which is then segmented with Discriminate Clustering and ends with an extraction process with the Gray Level Co-occurrence Matrix (GLCM) feature. The results will be classified into three groups [7]. Another study was to analyze the maturity level of mango fruit with neural networks and Computer Vision [8]. The network model used is the Backpropagation model that can detect the color of mango that is ripe or immature. Accuracy results obtained are 94% [8]. Determination of tomato maturity level with Backpropagation using lightness, redness or greenness, and yellowness or blueness colouring parameters denoted L *, a *, and b *. In addition to these three color features using nine color features extracted from these three parameters. The results of sorting on tomatoes were 97.9% [9].

Research on the determination of sugar content has been done. One of them is the prediction research of sugar content on citrus fruit using dummy variable method and correlation analysis obtained from image segmentation process [10]. By using the spectrum light on the image can be predicted sugar levels. The results will be grouped into three namely half sweet, sweet, and very sweet [10].

This research proposed application of regression method to predict sugar content in tomatoes based on RGB and HSV color models on tomato skin image because the HSV model represents more accurate image color information. In addition, this research was also proposed application of the

Backpropagation method to predict the level of maturity based on the red histogram on the tomato skin image.

II. RESEARCH METHODOLOGY

The stages of this study can be seen in Figure 1

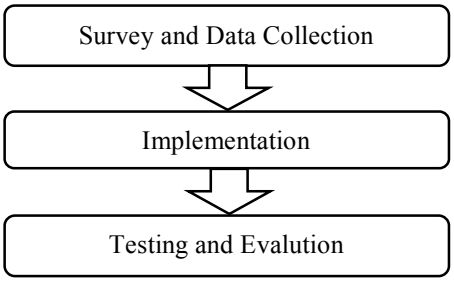


Figure 1 The Diagram of Research Stage

Each stages of research in Figure 1 is described as follows:

A. *Survey and Data Collection*

The tomato image data is groupe into three categories, namely : raw, half mature, and mature. Determination of tomato maturity has been validated by tomato traders. Each group has 30 samples consisting of 25 learning data and 5 testing data. So, the total is 75 tomatoes as training data and 15 tomatoes as testing data. Taking the image of tomato skin using the Lenovo S660 camera. Meanwhile, glucose data is taken using a refractometer.

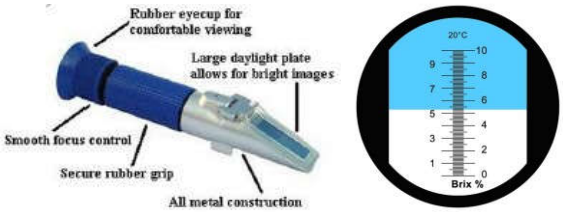


Figure 2 Specification of Image Capture Tool and Glucose Content Tool

The measuring scale of the refractometer can be seen in Figure 2. The measurement results are seen from white and blue light limits. As in figure 2, glucose levels are 5.2% Brix[11].

B. *Implementation*



Figure 3 The Diagram of Preprocessing

The preprocessing process on tomato skin image can be seen in Figure 3. Each chart in Figure 3 can be described as follows [12] :

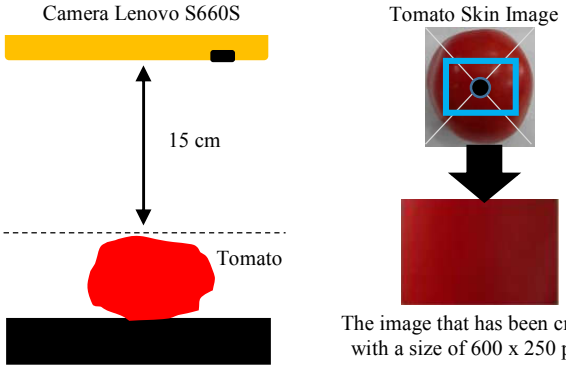


Figure 4 Illustration of Tomato Skin Image Retrieval(a) and Cropping Process on Tomato Skin Image(b)

a. *Image Acquisition*

This stage aims to improve image quality and facilitate the process of data identification. In this study, the tomato skin image was taken using Lenovo S660 camera with 8 MB specifications. The distance between the camera and the object is 15 cm. Illustration of taking tomato skin image data can be seen in Figure 4a. The first image will be cropped in the middle of the image with a size of 600 x 250 pixels. The second image will be taken on one side of the tomato skin which is the dominant it skin color. This image will be removed background. The croppping process can be seen in Figure 4b.

b. *Feature Extraction*

This stage specifies the red histogram, RGB color, and HSV color conversion. Determination of the red histogram is obtained from the image of tomato skin that has been cropped. Meanwhile, the calculation of RGB and HSV color values is obtained from the image of tomato skin that has been removed from the background. Determination of red histogram and RGB to HSV conversion can be written algorithm below this [16].

```
Determine Red Histogram
Input    : Image of tomato skin that has been cropped
Output   : Red Histogram
Process  :
    Begin
    Red_Histogram[256];
    RED[width*height];
    void Data_Histogram (File data) {
      image ← ImageIO.read(data);
    }
    int[ ] setRedImage () {
      k ← 0;
      for i ← 0 to height do {
        for j ← 0 to width do {
          Color c ← new Color(image.getRGB(j,i));
          RED[k] ← c.getRed();
          k++;
        endfor
      endfor
    return RED;
    }
    void setRedHistogram (int r) {
      Red_Histogram[r]++;
    }
    End
```

```
Conversion RGB to HSV
Input : RGB Model Color
Output        : HSV Model Color
Process       :
V ← Max(R,G,B)
Delta ← V-Min(R,G,B)
S ← choose(V ← 0, 0, Delta*255/V)
if S = 0
        H ← 0
else
        if R = V
                H ← (G-B)/Delta
        else if G = V
                H ← 2 + (B-R)/Delta
        else if B = V
                H ← 4 + (R-G)/Delta
        End
        if H < 0
                H += 6
        End
        H ← H*255/6
End
```

c. *Feature Matching*

This section will determine the level of maturity and predictions of glucose content in tomato based on vector feature of its image using backpropagation and linier regression. This step can be seen in Figure 5 and Figure 6.

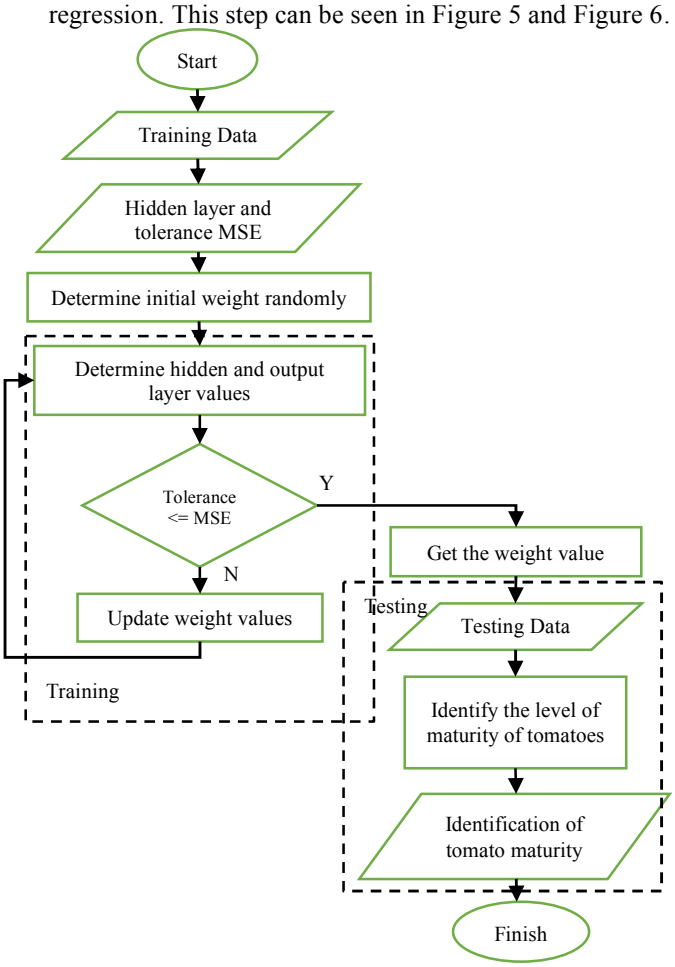


Figure 5 Diagram of Determining the Tomato Maturity Level

In Figure 5, determine hidden layer and output layer value can be calculated using equation as follow :

$$z_j = \frac{1}{1+e^{-(z_{net_j})}} \quad (1)$$

$$y_k = \frac{1}{1+e^{-(y_{net_k})}} \quad (2)$$

where $z_{net_j} = b_x + \sum_{i=0}^{n} x_i . v_{ij}$, x_i is input layer value in neuron to i, v_{ij} is the weight value input layer i to hidden layer j, $y_{net_k} = b_z + \sum_{j=0}^{n} z_j . w_{jk}$, z_j is hidden layer value in neuron to j, w_{jk} is the weight value hidden layer j to output layer k, y_k is output layer value in neuro to k, b_x is bias value on the input layer, and b_z is bias value on the hidden layer. Besides, calculate MSE (Mean Squard Error) values using formula :

$$MSE = \frac{1}{2}\sum_{k=1}^{N}(t_k - y_k)^2 \quad (3)$$

where t_k is target value in the output layer. This value is used to check whether it has reached the tolerance limit or not. If yes, then test the data. But, if not, must update the weight value. To update weight value can be calculated using equation as follow :

$$w_{jk_{new}} = w_{jk} + \Delta w_{jk} \quad (4)$$

$$v_{ij_{new}} = v_{ij} + \Delta v_{ij} \quad (5)$$

$$\Delta w_{jk} = \alpha . \delta_k . z_j \quad (6)$$

$$\Delta v_{ij} = \alpha . \delta_j . x_i \quad (7)$$

$$\delta_k = (t_k - y_k) . y_k . (1 - y_k) \quad (8)$$

$$\delta_j = \delta_{net_j} . z_j . (1 - z_j) \quad (9)$$

where $\delta_{net_j} = \sum_{k=1}^{m} \delta_k . w_{jk}$ and α is learning state value. After that, repeat for determine hidden and output layer value until it reaches the tolerance limit.

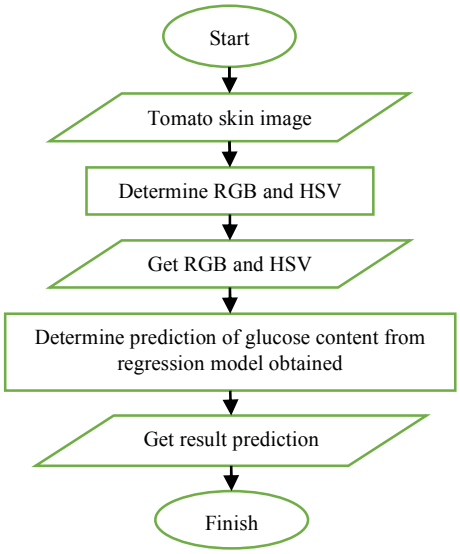


Figure 6 Diagram of prediction of Tomato Glucose Content

In Figure 6, to determine RGB and HSV can be calculated using formula as follow :

$$\bar{x} = \frac{\sum_{i=0}^{L_{tomato}} x_i}{L_{tomato}} \quad (10)$$

where $\bar{x}$ is average color values, L_{tomato} is area of tomato, and x_i is pixel color values. this value will be used to determine the correlation value and create a regression equation to calculate the predicted glucose content.

C. *Testing and Evaluation*

The predicted results of glucose levels and maturity levels of tomatoes will be determined accurately. Calculation of accuracy on the prediction of glucose levels using abstract error. Meanwhile, the maturity level is determined by confusion matrix.

To calculate the error value in the predicted tomato sugar content can be calculated by the equation below: [13]

$$\%Error = \left|\frac{G_e - G_p}{G_e}\right| \times 100\% \quad (11)$$

where G_e is the glucose content value based on a reflactometer and G_p is the glucose content value based on the calculation of the regression equation.

Calculation of accuracy in determining the maturity level of tomatoes is applied using a confusion matrix. This calculation is seen from many classifications [14]. The level of accuracy with this method can be seen in the value of accuracy, precision, and recall. These three values can be calculated as follows [15] :

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (12)$$

$$precision = \frac{TP}{TP+FP} \quad (13)$$

$$recall = \frac{TP}{TP+FN} \quad (14)$$

which TP is True Positive, TN is True Negative, FP is False Positive, and FN is False Negative [15].

III. RESULT

A. Determination of Tomato Maturity

The maturity level of tomatoes was determined using a red histogram on the image of tomato skin that has been cropped. This value will be used as an input layer in the backpropagation network. Then, obtained three output layer values to identify the level of maturity. The form of the program of determining the level of tomato maturity can be seen in Figure 7.

TABLE I. TEST RESULT OF TOMATO MATURITY

MSE	Confusion Matrix	Hidden Layer				
		1	2	3	4	5
0.1	*Accuracy*	41.33%	73.33%	68%	68%	65.33%
	Precision	30.62%	75.41%	62.59%	65.31%	56.13%
	Recall	41.33%	73.33%	61.4%	68%	65.33%
0.01	*Accuracy*	56%	66.66%	80.02%	86.67%	72%
	Precision	43.56%	72.45%	83.04%	90.48%	71.43%
	Recall	56%	66.66%	80.02%	86.67%	72%
0.001	*Accuracy*	48%	57.33%	66.67%	66.67%	66.67%
	Precision	42.95%	59.35%	64.44%	64.44%	64.44%
	Recall	48%	57.33%	66.67%	66.67%	66.67%

Before identifying tomato maturity, first do the learning process with 75 images of tomato skin as exercise data with the number of hidden layer neurons and specified MSE limits. The learning process will stop if the MSE value has reached the tolerance limit. If not, it will update each weight on the backpropagation network.. Next, testing the data will be done 5 times for each hidden screen neuron. Many hidden screen neuron that are used is 1 to 5 neurons. This test will get accuracy, precision, and recall.

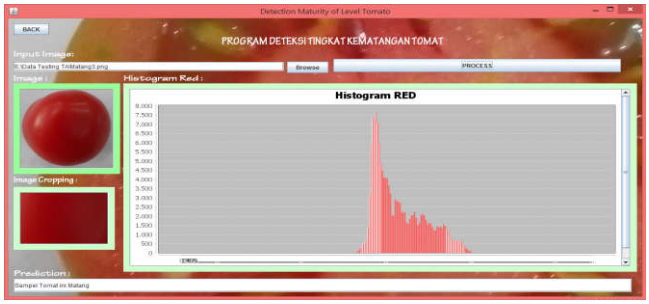


Figure 7 Interface of Determine of Tomato Maturity

TABLE II. TEST RESULT OF TOMATO DATA (WITH 4 HIDDEN LAYER & MSE = 0,01)

Sample	Test to-				
	1	2	3	4	5
1	Raw	Raw	Raw	Raw	Raw
2	Raw	Raw	Raw	Raw	Raw
3	Raw	Raw	Raw	Raw	Raw
4	Raw	Raw	Raw	Raw	Raw
5	Raw	Raw	Raw	Raw	Raw
6	Raw	Raw	Raw	Raw	Raw
7	Half-Mature	Half-Mature	Half-Mature	Half-Mature	Half-Mature
8	Half-Mature	Half-Mature	Half-Mature	Half-Mature	Half-Mature
9	Raw	Raw	Raw	Raw	Raw
10	Half-Mature	Half-Mature	Half-Mature	Half-Mature	Half-Mature
11	Mature	Mature	Mature	Mature	Mature
12	Mature	Mature	Mature	Mature	Mature
13	Mature	Mature	Mature	Mature	Mature
14	Mature	Mature	Mature	Mature	Mature
15	Mature	Mature	Mature	Mature	Mature

The MSE values used for tolerance limits during the learning process are 0.1, 0.01, and 0.001 with a limit of 5000000 epochs. After that, test the data with 15 tomato skin images as testing data.. This results can be obtained in Table II. In Table II, the best accuracy is testing with 4 neuron hidden layers and MSE = 0.01 with accuracy, precision, and recall values of 86.67%, 90.48%, and 86.67 %. That results can be seen in Table I. In Table I, the accuracy obtained on raw, half-mature, and mature classes is 100%, 60%, and 100%.

B. Prediction of Tomato Glucose Content

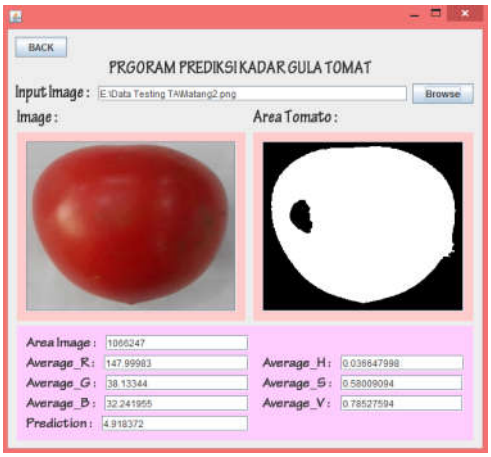


Figure 8 Interface of Prediction of Glucose Content

Predicted glucose content of tomatoes is seen based on the average value of RGB and HSV. The prediction program for glucose content in tomatoes shows the average values on RGB and HSV colors, tomato area and predictive value of glucose content. The form of this program can be seen in Figure 8.

Each variable R, G, B, H, S, and V will determine the correlation. Correlation calculation results can use the equation below [19] :

$$r = \frac{n\sum_{i=1}^{n} X_i.Y_i - (\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{\sqrt{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}.\sqrt{n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}} \quad (15)$$

Let the calculation of the correlation between the color of Hue (X) with glucose content (Y) in the raw class in Table III. In Table III, we can calculated $\sum_{i=1}^{n} X_i = 4.943$, $\sum_{i=1}^{n} Y_i = 98.4$, $\sum_{i=1}^{n} X_i^2 = 0.989273$, $\sum_{i=1}^{n} X_i.Y_i = 19.7624$, dan $\sum_{i=1}^{n} Y_i^2 = 401.6$ with $n = 25$. The results will be obtained in the calculation below :

TABLE III. DATA BETWEEN HUE COLOR WITH GLUCOSA CONTENT ON RAW CLASS

Sampel	X	Y	Sampel	X	Y
1	0.201	3.5	14	0.183	3.8
2	0.198	3.4	15	0.19	3.8
3	0.172	3	16	0.209	5.2
4	0.187	2.9	17	0.242	4.5
5	0.182	3.2	18	0.211	5.2
6	0.164	3	19	0.241	4.8
7	0.187	2.9	20	0.205	4
8	0.18	2.9	21	0.217	4.9
9	0.173	3.8	22	0.218	4.8
10	0.186	3.2	23	0.205	4.2
11	0.172	3.9	24	0.227	5
12	0.182	3.7	25	0.232	4.6
13	0.179	4.2			

$$r = \frac{(25)(19.7624)-(4.943)(98.4)}{\sqrt{(25)(0.989273)-(4.9423)^2}\sqrt{(25)(401.6)-(98.4)^2}}$$

$$r = \frac{7.6688}{(0.546421)(18.90608)}$$

$$r = 0.742$$

The result of correlation calculation for some other color variables can be seen in Table IV.

TABLE IV. CORRELATION VALUE DATA ON LINEAR REGRESSION

Variable	Raw	Half - Mature	Mature	All
R	-0.549	0.698	0.798	0.532
G	0.003	0.16	0.271	-0.449
B	-0.195	0.464	0.355	-0.172
R+B	-0.479	0.723	0.802	0.493
| R-G |	0.668	0.616	0.651	0.621
$\frac{R+B}{G+B}$	-0.676	0.588	0.163	0.502
$\frac{\lvert R-G\rvert}{R+B}$	0.688	0.545	0.14	0.619
$\frac{R}{R+B}$	0.535	-0.69	-0.799	-0.414
$\frac{G}{R}$	0.681	-0.59	-0.055	-0.455
H	0.742	-0.623	-0.567	-0.452
S	0.099	0.066	0.391	0.614
V	0.015	0.709	0.798	0.614
H+S	0.344	-0.379	-0.47	0.003
H+V	0.627	0.623	0.042	0.383
S+V	0.097	0.551	0.747	0.602
$\frac{S}{S+V}$	0.089	-0.74	-0.783	-0.444
$\frac{V}{H+S}$	-0.25	0.756	0.761	0.551
$\frac{V}{S+V}$	-0.089	0.74	0.783	0.445
$\frac{R}{H}$	-0.742	0.584	0.419	0.48
$\frac{H+S}{S+V}$	0.509	-0.723	-0.675	-0.529
$\frac{R}{V}$	-0.012	-0.69	-0.793	-0.581
$\frac{R}{S+V}$	-0.116	-0.552	-0.752	-0.591

Determination of regression equations for prediction of glucose content using two variables from Table IV. The two variables are taken from each class with the highest correlation value. In the raw class, the two variables with the greatest correlation value is H and $\frac{|R-G|}{R+B}$. Neither with any other classes. In the half-baked class, the variables used are $\frac{V}{H+S}$ and $\frac{V}{S+V}$. In the mature class, the variables used are $R+B$ and R. Meanwhile, the whole class using variable $|R-G|$ and $\frac{|R-G|}{R+B}$.

The model form of multiple regression equations with two variables can be written in equation [17] :

$$Y = a + b_1X + b_2Z \qquad (16)$$

Parameters a, b_1, and b_2, can be found by equations (7), (8), and (9) [18].

$$\sum_{i=1}^{n} Y_i = n.a + b_1.\sum_{i=1}^{n} X_i + b_2.\sum_{i=1}^{n} Z_i \qquad (17)$$

$$\sum_{i=1}^{n} X_i.Y_i = a.\sum_{i=1}^{n} X_i + b_1.\sum_{i=1}^{n} X_i^2 + b_2.\sum_{i=1}^{n} X_i.Z_i \qquad (18)$$

$$\sum_{i=1}^{n} Z_i.Y_i = a.\sum_{i=1}^{n} Z_i + b_1.\sum_{i=1}^{n} X_i.Z_i + b_2.\sum_{i=1}^{n} Z_i^2 \qquad (19)$$

Suppose the determination of glucose content with variable H and $\frac{|R-G|}{R+B}$ in raw class in Table V. With X = Hue color , Y = glucose content, dan $Z = \frac{|R-G|}{R+B}$. In Table V, we can calculated $\sum_{i=1}^{n} X_i.Z_i = 0.565$, $\sum_{i=1}^{n} Z_i.Y_i = 11.426$, dan $\sum_{i=1}^{n} Z_i^2 = 0.424$ then the results obtained are the three equations below (in matrix form) :

TABLE V. DATA BETWEEN H AND $\frac{|R-G|}{R+B}$ WITH GLUCOSE CONTENT ON RAW CLASS

Sampel	X	Z	Y	Sampel	X	Z	Y
1	0.201	0.149	3.5	14	0.183	0.059	3.8
2	0.198	0.117	3.4	15	0.19	0.088	3.8
3	0.172	0.041	3	16	0.209	0.134	5.2
4	0.187	0.074	2.9	17	0.242	0.238	4.5
5	0.182	0.044	3.2	18	0.211	0.173	5.2
6	0.164	-0.007	3	19	0.241	0.239	4.8
7	0.187	0.059	2.9	20	0.205	0.126	4
8	0.18	0.084	2.9	21	0.217	0.165	4.9
9	0.173	0.038	3.8	22	0.218	0.181	4.8
10	0.186	0.059	3.2	23	0.205	0.124	4.2
11	0.172	0.013	3.9	24	0.227	0.223	5
12	0.182	0.012	3.7	25	0.232	0.221	4.6
13	0.179	-0.001	4.2				

$$\begin{pmatrix} 25 & 4.943 & 2.653 \\ 4.943 & 0.989 & 0.564 \\ 2.653 & 0.564 & 0.424 \end{pmatrix} \begin{pmatrix} a \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} 98.4 \\ 19.762 \\ 11.426 \end{pmatrix}$$

The result of these three parameters is -3.95373, 42.6213, and -5.06267, respectively. Then, the regression equation model obtained for the prediction of glucose content in the crude class represented by YR is

$$YR = -3{,}95372 + 42{,}6213H - 5{,}062668\left(\frac{|R-G|}{R+B}\right)$$

The determination of the regression equation model in the half-Mature, mature, and combined classes of the three classes is done as in the previous technique. Thus, the model of the regression equation for the raw class (YR), half-baked (YHC), mature (YM), and the combination of the three classes (Y) is obtained:

$$YR = -3.95372 + 42.6213H - 5.062668\left(\frac{|R-G|}{R+B}\right)$$

$$YHC = 3.2733 + 4.5602\left(\frac{V}{H+S}\right) - 3.2996\left(\frac{V}{S+V}\right)$$

$$YM = 1.340416 + 0.012940(R+B) + 0.0084620R$$

$$Y = 3.82505 + 0.00598(|R-G|) + 0.715703\left(\frac{|R-G|}{R+B}\right)$$

The four models of regression equations can be described in a 3-dimensional graphic form using Curve Expert. The results are obtained in Figure 9.

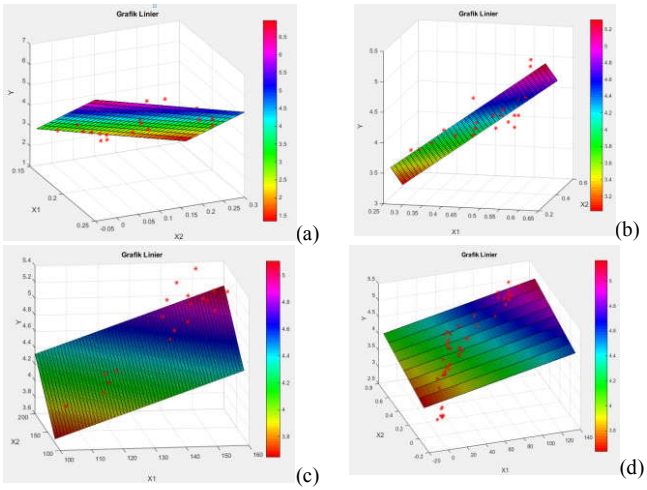


Figure 9 Linear Regression Model of Two Variables In Raw (a), Half Matured (b), Mature (c), and Three-Grade Class (d)

TABLE VI. TEST RESULT ON RAW CLASS

Sampel	G_e	H	$\frac{\lvert R-G\rvert}{R+B}$	G_p	%Error
1	4.2	0.194	0.094	3.837	8.649
2	4.3	0.23	0.236	4.657	8.298
3	4.3	0.189	0.089	3.652	15.07
4	4.2	0.229	0.23	4.641	10.49
5	4.3	0.178	0.053	3.365	21.74
Average Error					12.85

From the regression equation model obtained will be tested and calculated the average value of error. The determination of error can be seen in Table VI up to Table IX. For example, in sample 1 of raw skin tomato skin image, with $H = 0.194$ and $\frac{|R-G|}{R+B} = 0.094$, the predicted glucose content obtained is

$G_p = -3.95372 + 42.6213(0.194) - 5.062668(0.094)$

$G_P = -3.95372 + 8.268 - 0.4758$

$G_p = 4.214$

After that, calculate the error value, so it is obtained

$\%Error = \left|\frac{4.214-4.9}{4.9}\right| \times 100\%$

$\%Error = 14.01\%$

TABLE VII. TEST RESULT ON HALF-MATURE CLASS

Sampel	G_e	$\frac{V}{H+S}$	$\frac{V}{S+V}$	G_p	%Error
1	4.9	0.457	0.347	4.214	14.01
2	4.2	0.548	0.383	4.51	7.384
3	4	0.563	0.392	4.549	13.73
4	4.9	0.454	0.345	4.204	14.2
5	4.2	0.551	0.382	4.524	7.704
Average Error					11.41

TABLE VIII. TEST RESULT ON MATURE CLASS

Sampel	G_e	R+B	R	G_p	%Error
1	5	161.4	131.3	4.54	9.196
2	4.4	180.2	148	4.925	11.94
3	4.4	169.1	139.9	4.712	7.1
4	5.2	166.8	135.7	4.647	10.64
5	4.6	176.7	145.3	4.856	5.559
Average Error					8.89

Neither is the half-mature, mature, and combined class of three classes. The results of the tests on crude, half-Mature, mature, and three-joint classes can be seen in the table in order, Table VI, VII, VIII, and IX.

TABLE I. TEST RESULT ON COMBINATION OF THREE CLASS (Raw, Half-Mature, and Mature)

Sampel	G_e	| R-G |	$\frac{\lvert R-G\rvert}{R+B}$	G_p	%Error
1	4.2	10.94	0.094	3.958	5.761
2	4.3	22.26	0.236	4.127	4.029
3	4.3	10.21	0.089	3.95	8.147
4	4.2	22.49	0.23	4.124	1.8
5	4.3	6.556	0.053	3.902	9.254
6	4.9	20.32	0.156	4.059	17.17
7	4.2	35.72	0.242	4.212	0.292
8	4	37.49	0.245	4.224	5.608
9	4.9	20.65	0.163	4.065	17.03
10	4.2	41.63	0.284	4.277	1.839
11	5	98.78	0.612	4.854	2.924
12	4.4	109.9	0.61	4.918	11.78
13	4.4	109.7	0.649	4.945	12.39
14	5.2	86.38	0.518	4.712	9.376
15	4.6	105.7	0.599	4.886	6.214
Average Error					7,574

The result of the test in the raw class obtained the average value of the error of 12.85%. While the average value of error for the half-Mature and mature class is 11.41% and 8.89%. If the three categories are combined, then the test results obtained by 7.57%. The average error results obtained from Tables VI, VII, and VIII when averaged will be obtained at 11.05%. If the result is compared to the average error rate in Table IX, then the result is better. Thus, the prediction of glucose content with the Y equation is better than the YR, YHC, and YM equations.

IV. CONCLUSION

Based on the results of the study can predict glucose levels and determine the level of maturity of tomatoes. Prediction of glucose level has error value equal to 7.57% with model equation $Y = 3.825 + 0.00598(|R - G|) + 0.715703\left(\frac{|R-G|}{R+B}\right)$. Meanwhile, determination of maturity level with network Backpropagation has an accuracy of 86.67% with four hidden layers and a 0.01 MSE limit with raw, half-baked, and mature 100%, 60%, and 100% accuracy.

REFERENCES

[1] S. Melesse, N. Sobratee, & T. Workneh. (2016). Application of logistic regression statistical technique to evaluate tomato quality subjected to different pre-and post-harvest treatments. *Biological Agriculture & Horticulture*, *32*(4), 277-287.

[2] Y. Supriyati, Siregar, & D. Firmansyah. (2009). *Bertanam Tomat Dalam Pot dan Polibag*. Penerbit : Penebar Swadaya. Depok.

[3] BPS. (2017). Konsumsi Buah dan Sayur Susenas Maret 2016

[4] K. Warman, L. A. Harahap, & A. P. Munir. (2015). Identification of Citrus Maturity with Artificial Neural Network. *Jurnal Rekayasa Pangan dan Pertanian*, *3*(2), 248-25.

[5] Y. Permadi. (2015). Aplikasi Pengolahan Citra Untuk Identifikasi Kematangan Mentimun Berdasarkan Tekstur Kulit Buah Menggunakan Metode Ekstraksi Ciri Statistik. *Jurnal Informatika*, *9*(1).

[6] U. K. Acharya, P. P. Subedi, & K. B. Walsh. (2017). Robustness of Tomato Quality Evaluation Using a Portable Vis-SWNIRS for Dry Matter and Colour. *International journal of analytical chemistry*, *2017*.

[7] Sagayaraj, A. Stephen., C. Kalaiarasan, S. Gokulnath, R. Karthikeyan, & K. Bharathi. (2017). Detection of Maturity Level and Disease in Tomato Using Discriminative Clustering Method. *International Journal of Innovative Research in Computer and Communication Engineering*. *ISO 3297*. 4771-4779.

[8] E. H. Yossy, J. Pranata, T. Wijaya, H. Hermawan, & W. Budiharto. (2017). Mango Fruit Sortation System using Neural Network and Computer Vision. *Procedia Computer Science*, *116*, 596-603.

[9] S. Kassem, M. A. Sabbah, A. M. Aboukarima, & R. M. Kamel. (2015). A study on color sorting of tomatoes maturity using machine vision and artificial neural networks. *Egyptian Journal of Agricultural Research*, *93*(1), 147-161.

[10] X. Wang, C. Wu, & M. Hirafuji. (2016). Visible Light Image-Based Method for Sugar Content Classification of Citrus. *PloS one*, *11*(1), e0147419.

[11] A. A. Bhosale. (2017). Detection of Sugar Content in Citrus Fruits by Capacitance Method. *Procedia Engineering*, *181*, 466-471.

[12] N. Astriananda, & F. S. Mohamad. (2017). Ripeness Identification of Tomato Using Different Color Models Based on Neural Networklevenberg-Marquardt

[13] S. C. Chapra, & R. P. Canale. (1998). *Numerical methods for engineers* (Vol. 2). New York: McGraw-Hill.

[14] S. Visa, B. Ramsay, A. L. Ralescu, & E. Van Der Knaap. (2011, April). Confusion Matrix-based Feature Selection. In *MAICS* (pp. 120-127).

[15] T. R. Patil, & S. S. Sherekar. (2013). Performance analysis of Naive Bayes and J48 classification algorithm for data classification. *International Journal of Computer Science and Applications*, *6*(2), 256-261.

[16] P. M. Nishad. (2013). Various colour spaces and colour space conversion. *International Journal of Global Research in Computer Science (UGC Approved Journal)*, *4*(1), 44-48.

[17] R. E. Walpole, R. H. Myers, S. L. Myers, & K. Ye. (2014). *Probability and statistics for engineers and scientists*. London: Pearson.

[18] L. S. Atmaja. (2009). Statistika untuk bisnis dan ekonomi. *Penerbit Andi, Yogyakarta*.

[19] E. Heron. (2009). Correlation and regression. *Retrieved July 21, 2014*.

Conformance Checking Evaluation of Process Discovery Using Modified Alpha++ Miner Algorithm

Yutika Amelia Effendi and Riyanarto Sarno
Department of Informatics, Faculty of Information and Communication Technology
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
email: yutika.effendi@gmail.com, riyanarto@if.its.ac.id

Abstract*—Process mining techniques extract business processes from event logs. Process discovery, one of process mining techniques, has aim to mine a process model from event log. To help its process, process discovery uses some algorithms such as alpha, alpha+ and alpha++ where each algorithm has some advantages and limitations in analyzing event log. Alpha++ miner algorithm is considered the most advance improvement of alpha and alpha+ algorithm, therefore we use alpha++ as basic algorithm to our Modified alpha++ Miner. The proposed method in this paper uses Modified alpha++ Miner to generate process model, determine sequential and parallel relation from process model using temporal causal relation and control flow, and then evaluate the process model in terms of fitness value, validity and completeness. This paper will be focused on the advantages of Modified alpha++ Miner to discover business process.

Keywords-- Alpha++ Algorithm; Time Interval; Alpha Miner; Process Discovery; Process Mining; Conformance Checking

I. INTRODUCTION

Process mining techniques extract business processes from event logs. Process mining is divided into three types, namely process discovery, model extension/ enhancement and conformance checking. From those three, mining a process model from event log is main focus of process discovery [1]. To analyze the event log, process discovery uses technique or algorithm, such as α (Alpha), α+, α++, and heuristics miner [2]. Each of this algorithm has some advantages compared to the others, but it also has its limitations [3]. Limitations of α, α+, and α++ algorithm can be found in related works part of this paper.

The proposed method in this paper uses Modified α++ Miner (MA++M) algorithm to generate the process model as well as their relation based on event log. Compared to α++, this algorithm can detect overlapping process in time interval and identifies it as a parallel process. Because of this advantage, MA++M algorithm will generate less trace than the α++ algorithm. We use α++ algorithm because at current time the α++ is the latest version of modified α algorithm with big improvement of either α or α+ and MTBAM algorithm is also based on α algorithm, short description related to α++ will be written in related works section of this paper.

Main focus of this research is modifying α++ algorithm to discover process model and then evaluate the process model in terms of fitness value, validity and completeness, which is better known as conformance checking.

II. RELATED WORK

Related work of this paper explains and does comparing the existing process discovery algorithm α++ based on the earlier α and α+ algorithm. α+ algorithm itself is based on the algorithm called α.

A. α++ Algorithm

α algorithm is the most commonly used algorithm in process discovery to analyze and discover activity from cases in the event log. However, α algorithm has some disadvantages in the implementation. In [4], the limitations of α algorithm is further exposed, α algorithm cannot mine duplicate and hidden tasks, non-free choice [5], cannot deal with noisy data, and data with time constraint and mining loops (L1L and L2L). α+ which is the improvement of α algorithm can detect short loop like length one loop. Lastly α++ algorithm that can detect implicit dependencies, discover length one loop (L1L) and length two loop (L2L) and can reconstruct workflow-net (WF-net) to non-free choice. However, there is a disadvantage of α++ algorithm, that is it will produce traces according to parallel cases in the event log. The bigger the number of the parallel case, the higher the number of traces generated from α++ algorithm [6]. Further explanation of α++ algorithm and α algorithm can be found in [7].

B. Modified Time-Based Alpha Miner Algorithm (MTBAM)

MTBAM algorithm uses temporal causal relation and control flow to differentiate sequential and parallel relation as explained in [6]. As mentioned before in introduction, this algorithm can generate less traces than α algorithm using time interval so the fitness value is higher than α algorithm. We adopt this temporal causal relation and control flow to determine the parallel and sequence of process model based on their time interval into this modified α++ miner algorithm.

III. CONFORMANCE CHECKING EVALUATION

In this section, conformance checking evaluation is explained. We only use three evaluation criteria in this research. They are fitness value, validity and completeness. Process model from discovery process using Modified α++ Miner algorithm are evaluated based on this criteria.

A. Fitness of Process Model

Conformance checking is closely related to fitness. Fitness is kind of evaluation of a process model which has a focus on all activities of process model should correctly parsed based on event log. The process model which has all the activities that can be correctly parsed based on the event log has a high fitness value. On the contrary if many of the activities cannot be correctly parsed into the process model then the fitness value will be low.

Fitness value ranges from 0 (no activities including cases and traces parsed based on event log into process model) to 1 (all activities including cases and traces parsed correctly based on event log into process model) [6 , 8]. Generally, all α, α+ and α++ algorithm have high fitness value because of the relations in process model. They cover all relations, sequence and parallel without giving the threshold. We can calculate the fitness value using (1).

$$Xf = \frac{CapturedCasesinEL}{LogCasesinEL} \quad (1)$$

where:

Xf	: fitness value of process model
CapturedCasesinEL	: total number of cases that can be parsed into the process model
LogCasesinEL	: total number of cases in event log

B. Validity of Process Model

Event log that has been formed into a process model must be evaluated in terms of dependency graph, causal net and semantic. If all of them are in correct way, then it is meaning of validity [1 , 9]. Semantic process model can be obtained if process model enriches with split and join. To fulfil the evaluation criteria of validity, there are three steps to evaluate the validity of process model:

1. Determine dependency graph
2. Transform (1) and generate causal net
3. Create semantic process model
4. Obtain final model of event log which follows the validity condition correctly

C. Completeness of Process Model

Completeness is a notion referring to a problem in computational complexity theory. In process mining, concept of completeness related to minimum number of traces needed to discover process model correctly [10].

Process discovery also implement the idea of completeness. Process discovery works correctly when the event log meets the idea of the completeness of the algorithm. Event log which do not meet the idea of completeness are called incompleteness. Incompleteness occurs because there is too little data on event log which leads to misunderstandings in process discovery. So the result of process model is not as it should be [11].

Definition 1. There are event log (EL), process model (PM), and activity (Act). Notion of completeness for Act, EL and PM are:

$$Act \in PM, iff\ Act \in EL$$
$$(Act_1 \rightarrow Act_2) \in PM, iff\ (Act_1 > Act_2) \in EL\ where\ i\ in\ Act_i\ is\ an\ identifier\ of\ an\ activity.$$
$$(Act_1 \parallel Act_2) \in PM, iff\ [(Act_1 > Act_2) \in EL\ and\ (Act_2 > Act_1) \notin EL]\ \vee$$
$$[\{Act_1 @ Act_2 \vee Act_{1f} Act_2 \vee Act_1 \Diamond Act_2 \vee Act_{1p} Act_2 \vee Act_1 \square Act_2\} \in EL\ ].$$

Given a process model (PM), the number of activities (n) in the i-branch (i), and the number of activities which has branch activities (p). Process model (PM) consists of parallel AND, OR and XOR. The number of traces required for the process:

Formula 1. Parallel AND relation

$$AND_T = \sum_{i=1}^{p} \sum_{j=i+1}^{p} n_i \times n_j \quad (2)$$

Formula 2. Parallel OR relation

$$OR_T = \left(\sum_{i=1}^{p} \sum_{j=i+1}^{p} n_i \times n_j \right) + 1 \quad (3)$$

Formula 3. Parallel XOR relation

$$XOR_T = \sum_{i=1}^{p} if\ R_i == sequence, 1\ else\ 0 \quad (4)$$

Formula 4. Process model (PM) consists of a parallel set activity which each branch of parallel set has another parallel or sequence relation. To mine the process model with parallel set (P_T), the number of traces required:

$$P_T = AND_T + OR_T + XOR_T \quad (5)$$

IV. PROPOSED METHOD

Fig. 1 explains our proposed method of this research. This research focuses on discovering process model from Modified α++ miner algorithm. After we discover the process model, we evaluate the fitness, validity and completeness.

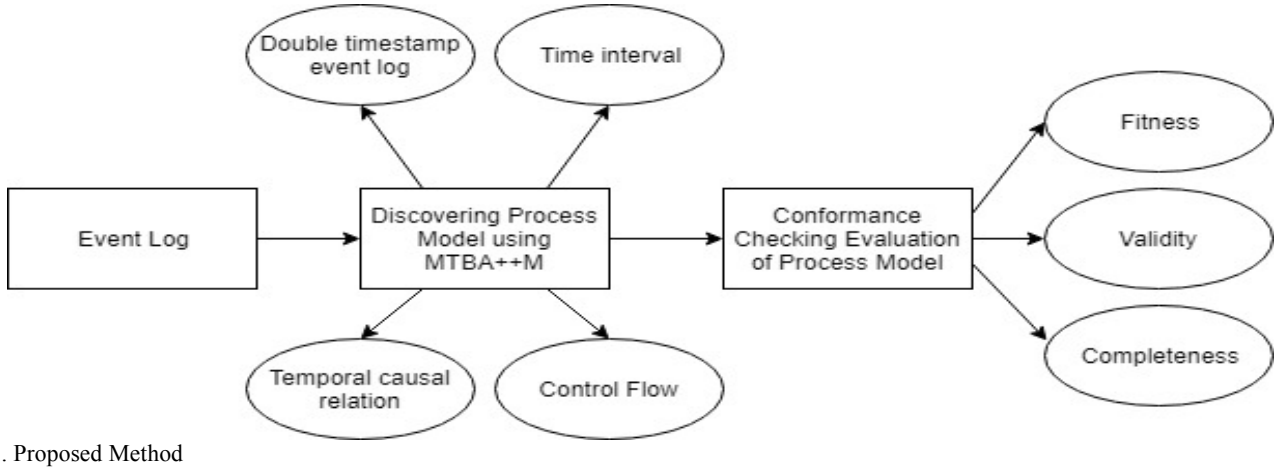


Fig. 1. Proposed Method

Step 1. Input data of this process discovery is event log. Event log has two types in reality; single timestamp and double timestamp event log. Generally, to discover the process model, we use Case ID, Activity and timestamp.

Step 2. After the input is ready to mine, Modified α++ Miner algorithm is used to discover the process model. To discover process model, as we explained in Section II, we adopt the steps of Modified Time-based Alpha Miner (MTBAM) algorithm to determine the sequence and parallel relation of process model which are temporal causal relation and control flow [6] [12] [13] [14].

Main differences between MTBAM and MA++M are in terms of detecting implicit dependency, discovering length one loop (L1L) & length two loop (L2L) and reconstruct the workflow-net (WF-net) into non-free choice.

There are 13 steps used in Modified α++ Miner algorithm:

Step 1. Determine the sequence relation and parallel relation from each case

Step 2. Delete duplicate relations (sequence and parallel) from all cases

Step 3. Get all traces of event log which include sequence and parallel relations

Step 4. Create set of transition in WF-net

Step 5. Create set of input and output transition from event log

Step 6. Create set of L1L and L2L from event log

Step 7. Detect all implicit dependencies from WF-net

Step 8. Create gantt chart for all traces

Step 9. Create the places

Step 10. Discover the process model

Step 11. Determine the type of parallel relation AND, XOR or OR

Step 12. Add input activity, output activity and sequence relations into process model

Step 13. A process model is complete

Step 3. Conformance Checking Evaluation of Process Model. After process model is generated, fitness value, validity and completeness of process model need to be measured.

V. EXPERIMENTAL RESULT

A. Event Log

We use Yarn Manufacturing Process event log in this research. Yarn Manufacturing Process consists of 14 activities, 8 traces and 50 cases. This event log is double timestamp event log which has start time and complete time. The relations in this event log are sequential and parallel relations. Table I shows event log of Yarn Manufacturing Process. For this experiment, we only use case id, activities and timestamp.

B. Experimental Result

We mine the process model using MA++M algorithm and event log in Table I. Fig. 2 shows the result after MA++M is run. Process model in Fig. 2 contains sequential relation, parallel XOR and OR, and also L1L and L2L. Table II explains activities listed in the process model in Fig. 2.

Based on Fig. 2, activity 'OpposingSpike' and activity 'AirCurrentBlowing' are in parallel relation XOR. Meanwhile, activity 'DrawingFrame' and activity 'RovingFrame' are in parallel relation OR. We get the knowledge that using MA++M, we can discover the process model which contains L1L, L2L and non-free choice. α, α+ and α++ algorithm cannot handle those three problems.

α algorithm can only discover relations of process model correctly. Fig. 3 shows process model discovery using α algorithm. Based on Fig. 2 and Fig.3, Modified α++ Miner algorithm is more powerful than α, α+ and α++ algorithm because all the conditions in the event log can be handled well. Activities listed in the process model in Fig. 3 also as same as explained in Table II.

TABLE I. EVENT LOG OF YARN MANUFACTURING PROCESS

Case ID	Activity	Start Time	End Time
PP1	Sending good receive	20/06/2014 08:32	20/06/2014 13:42
PP1	Getting good receive	20/06/2014 13:42	20/06/2014 23:41
PP1	Bale opening	20/06/2014 23:41	21/06/2014 08:16
PP1	Conditioning of MMP Fiber	21/06/2014 08:16	21/06/2014 10:46
PP1	Blending	21/06/2014 10:46	21/06/2014 16:57
PP1	Opposing spike	21/06/2014 16:57	21/06/2014 18:09
PP1	Striking cotton	21/06/2014 18:09	21/06/2014 19:15
PP1	Carding	21/06/2014 19:15	22/06/2014 10:51
PP1	Roving frame	22/06/2014 10:51	22/06/2014 16:28
PP1	Drawing frame	22/06/2014 16:28	22/06/2014 23:42
PP1	Combing	22/06/2014 23:42	23/06/2014 04:48
PP1	Ring framing	23/06/2014 04:48	23/06/2014 16:44
PP1	Cone winding	23/06/2014 16:44	23/06/2014 23:26
PP2	Sending good receive	23/06/2014 23:26	24/06/2014 04:19
PP2	Getting good receive	24/06/2014 04:19	24/06/2014 07:48
PP2	Bale opening	24/06/2014 07:48	25/06/2014 01:08
PP2	Conditioning of MMP Fiber	25/06/2014 01:08	25/06/2014 03:02
PP2	Blending	25/06/2014 03:02	25/06/2014 05:11
PP2	Air current blowing	25/06/2014 05:11	25/06/2014 08:25
PP2	Striking cotton	25/06/2014 08:25	25/06/2014 12:45
PP2	Carding	25/06/2014 12:45	26/06/2014 00:12
PP2	Drawing frame	26/06/2014 00:12	26/06/2014 04:48
PP2	Combing	26/06/2014 04:48	26/06/2014 15:16
PP2	Ring framing	26/06/2014 15:16	26/06/2014 22:49
PP2	Cone winding	26/06/2014 22:49	27/06/2014 05:19
PP3	Sending good receive	27/06/2014 05:19	27/06/2014 11:40
PP3	Getting good receive	27/06/2014 11:40	27/06/2014 20:00
PP3	Bale opening	27/06/2014 20:00	28/06/2014 07:10

TABLE II. CODE AND ACTIVITY NAME OF YARN MANUFACTURING PROCESS

Code	Activity Name
A	SendingGoodReceive
B	GettingGoodReceive
C	BaleOpening
D	ConditioningofMMFFiber
E	Blending
F	OpposingSpike
G	AirCurrentBlowing
H	StrikingCotton
I	Carding
J	DrawingFrame
K	RovingFrame
L	Combing
M	RingFraming
N	ConeWinding

C. Conformance Checking Evaluation

After discovering process model, we do the conformance checking evaluation of process model discovered by MA++M algorithm. We implement the (1), (2), (3), (4) and (5).

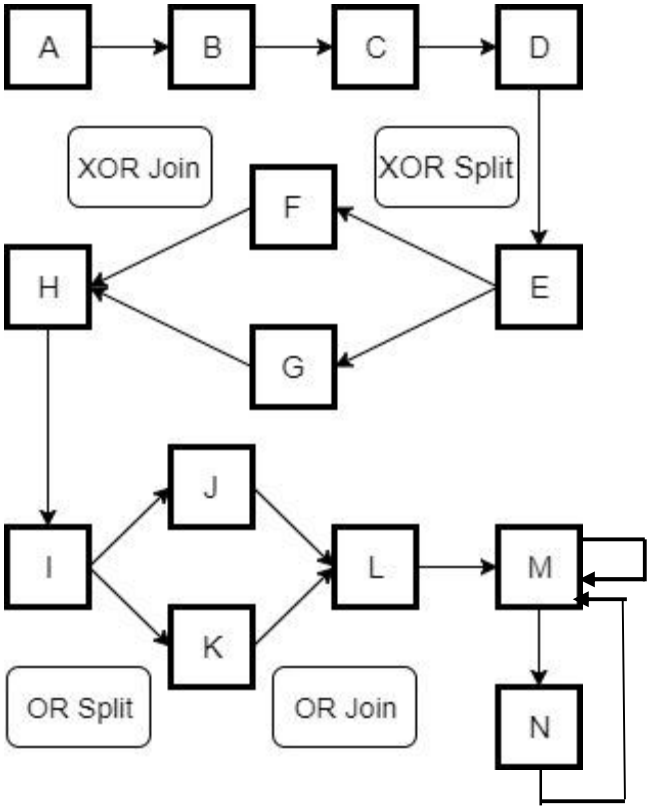


Fig. 2. Process Model of Yarn Manufacturing Process using MA++M algorithm

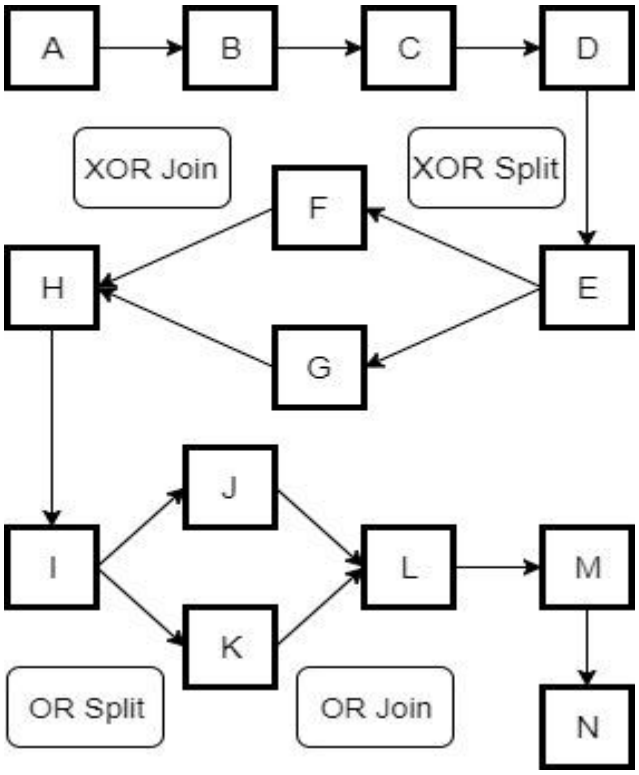


Fig. 3. Process Model of Yarn Manufacturing Process using α algorithm

a. Fitness

First evaluation is calculating the fitness value. We calculate fitness using (1) and we also compare the fitness between MA++M algorithm and MTBAM algorithm.

Based on Table III, result of fitness value of MA++M is higher than MTBAM because all cases in Yarn Manufacturing Process parsed into the process model based on event log. Meanwhile, MTBAM cannot parse the case where there are L1L and L2L.

TABLE III. FITNESS VALUE OF MA++M AND MTBAM

Fitness value of MA++M algorithm	**Fitness value of MTBAM algorithm**
1.0	0.92857

b. Validity

To get the validity of process model, we need to obtain the dependency graph of process model as generated using Modified α++ Miner algorithm. As we get the dependency graph, we need to transform it into causal net. Table IV shows dependency graph of process model generated by MA++M algorithm. After obtaining the dependency graph, we transform the dependency graph into causal net as shown in Table V. In causal net, input activity and output activity must be correct for each activity.

The final step of validity is connecting the gateway parallel and sequence relation of process model correctly so that we can create the semantic process model. Fig.4 shows the result where process model has correct validity. Based on validity terms, the result of process model by Modified α++ Miner algorithm generated the semantic process model in a valid way.

c. Completeness

Completeness is measured by calculating all the relations of business process to obtain the total minimum traces. They are needed to mine the process model correctly. Event log of Yarn Manufacturing Process can fulfill the idea of completeness if it has two traces calculated using (5).

Process model in Fig. 4 has two parallel relations; XOR relation and OR relation. We calculate the completeness using formula of completeness for XOR and OR (3) and (4).

The conditional XOR relation in the process model has one activity in each of its branch, then the number of traces based on (4) is

$$XOR = 1 + 1 = 2$$

Meanwhile, conditional OR relation also has one activity in each branch, then the number of trace based on Formula 3 is

$$OR = (1 \times 1) + 1 = 2$$

Based on (5), the number of traces of Yarn Manufacturing Process which consist of parallel sets are selecting the highest number of traces of each parallel sets. In this case, the answer is two.

Based on the calculation, the event log of Yarn Manufacturing Process for process discovery has 8 traces and 50 cases. Therefore, the event log of Yarn Manufacturing Process fulfills the idea of completeness, because it contains 8 traces of the 2 traces it should have.

TABLE IV. DEPENDENCY OF YARN MANUFACTURING PROCESS

Input Set	**Activity**	**Output Set**
{∅}	SendingGoodReceive	{GettingGoodReceive}
{SendingGoodReceive}	GettingGoodReceive	{BaleOpening}
{GettingGoodReceive}	BaleOpening	{ConditioningofMMFFiber}
{BaleOpening}	ConditioningofMMFFiber	{Blending}
{ConditioningofMMFFiber}	Blending	{OpposingSpike}
{Blending}	OpposingSpike	{AirCurrentBlowing}
{Blending}	OpposingSpike	{StrikingCotton}
{OpposingSpike}	AirCurrentBlowing	{Carding}
{OpposingSpike}	StrikingCotton	{Carding}
{StrikingCotton}	Carding	{DrawingFrame, RovingFrame}
{Carding}	DrawingFrame	{Combing}
{Carding}	RovingFrame	{Combing}
{DrawingFrame, RovingFrame}	Combing	{RingFraming}
{Combing}	RingFraming	{ConeWinding}
{RingFraming}	ConeWinding	{∅}

VI. CONCLUSION

In this research, we propose the modification of α++ algorithm which is the extended version of α and α+ algorithm. The modification of α++ is known as Modified α++ Miner (MA++M) algorithm which uses time interval in discovering the process model. Our experimental results explained that the result of process model can discover the sequence, parallel, length one loop (L1L) and length two loop (L2L) correctly. After discovering process model, the result is then evaluated using conformance checking which consists of-

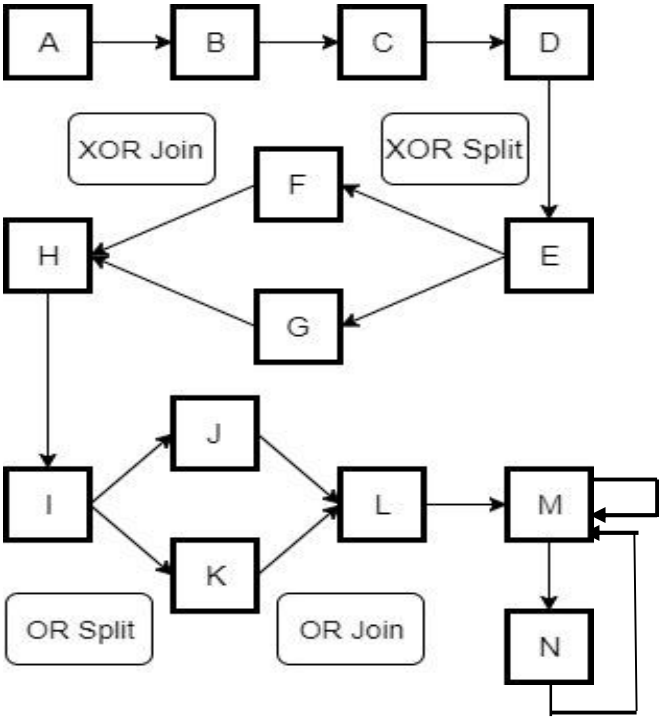


Fig. 4. Final process model of Yarn Manufacturing Process which has correct validity

three evaluation criteria; fitness value, validity and completeness. Final process model has high fitness value, can generate the semantics in a valid way and also fulfills the idea of completeness.

TABLE V. CAUSAL NET OF YARN MANUFACTURING PROCESS

Input Set	Activity	Output Set
{{∅}}	SendingGoodReceive	{{GettingGoodReceive}}
{{SendingGoodReceive}}	GettingGoodReceive	{{BaleOpening}}
{{GettingGoodReceive}}	BaleOpening	{{ConditioningofMMFFiber}}
{{BaleOpening}}	ConditioningofMMFFiber	{{Blending}}
{{ConditioningofMMFFiber}}	Blending	{{OpposingSpike}}
{{Blending}}	OpposingSpike	{{AirCurrentBlowing}}
{{Blending}}	OpposingSpike	{{StrikingCotton}}
{{OpposingSpike}}	AirCurrentBlowing	{{Carding}}
{{OpposingSpike}}	Striking_Cotton	{{Carding}}
{{StrikingCotton}}	Carding	{{DrawingFrame, RovingFrame}}
{{Carding}}	DrawingFrame	{{Combing}}
{{Carding}}	RovingFrame	{{Combing}}
{{DrawingFrame, RovingFrame}}	Combing	{{RingFraming}}
{{Combing}}	RingFraming	{{ConeWinding}}
{{RingFraming}}	ConeWinding	{{∅}}

REFERENCES

[1] R. Sarno, Y. A. Effendi, and F. Haryadita, "Modified Time-Based Heuristics Miner for Parallel Business Processes," *International Review on Computers and Software (IRECOS), vol. 11 (3), pp. 249-260, March 2016*. http://doi.org/10.15866/irecos.v11i3.8717

[2] R. Sarno and Y. A. Effendi, "Hierarchy Process Mining from Multi-Source Logs," *Telecommunication, Computing, Electronics and Control (TELKOMNIKA), Vol 15, No 4, 2017*. DOI: http://dx.doi.org/10.12928/telkomnika.v15i4.6326

[3] Banerjee and P. Gupta. "Extension of alpha algorithm for process mining"

[4] L. Wen, Wil M.P. van der Aalst, J. Wang, and J. Sun, "Mining process models with non-free-choice constructs".

[5] R. Sarno, P. L. I Sari, D. Sunaryono, B. Amaliah, and I. Mukhlash, "Mining decision to discover the relation of rules among decision points in a non-free choice construct," *Proceedings of International Conference on Information, Communication Technology and System (ICTS), 2014*. http://dx.doi.org/10.1109/ICTS.2014.7010557

[6] Y. A. Effendi and R. Sarno, "Modeling Parallel Business Process Using Modified Time-based Alpha Miner", *International Journal of Innovative Computing, Information and Control (IJICIC), Vol.14, No.4, August 2018*

[7] W.M.P. van der Aalst, Process mining: discovery, conformance and enhancement of business processes (Springer Science and Business Media, 2011). http://dx.doi.org/10.1007/978-3-642-19345-3

[8] Y. A. Effendi and R. Sarno, "Discovering optimized process model using rule discovery hybrid particle swarm optimization," *3rd International Conference on Science in Information Technology (ICSITech), pp. 97-103, 2017*. DOI: 10.1109/ICSITech.2017.8257092

[9] Y. A. Effendi and R. Sarno, "Discovering process model from event logs by considering overlapping rules," *4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), pp. 1-6, 2017*. DOI: 10.1109/EECSI.2017.8239193

[10] S. Huda, R. Sarno, and T. Ahmad, "Increasing Accuracy of Process-based Fraud Detection Using a Behavior Model," *International Journal of Software Engineering and Its Applications, 10 (5), pp. 175-188, 2016*. https://doi.org/10.14257/ijseia.2016.10.5.16.

[11] R. Sarno, W. A. Wibowo, D. Sunaryono, A. Munif. "Developing Workflow Patterns Based on Functional Subnets and Control–Flow Patterns". International Conference on Science in Information Technology (ICSITech), 2015. DOI : https://doi.org/10.1109/icsitech.2015.7407771

[12] S. S. Pinter and M. Golani, "Discovering workflow models from activities' life spans", *Computers in Industry, vol. 53, pp. 283-296, 2004*

[13] R. Sarno, Kartini, W. A. Wibowo, and A. Solichah, "Time Based Discovery of Parallel Business Processes," *International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, pp. *28 – 33, 2015*. DOI: 10.1109/IC3INA.2015.7377741

[14] I. G. Anugrah, R. Sarno, R. N. E. Anggraini, "Decomposition Using Refined Process Structure Tree (RPST) and Control Flow Complexity Metrics". *International Conference on Information & Communication Technology and Systems (ICTS),* 2015. DOI: https://doi.org/10.1109/icts.2015.7379899

Implementation of Bat Algorithm for COCOMO II Optimization

Yutika Amelia Effendi, Riyanarto Sarno, Joko Prasetyo
Department of Informatics, Faculty of Information and Communication Technology
Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
e-mail: yutika.effendi@gmail.com, riyanarto@if.its.ac.id, joko.prsty@gmail.com

***Abstract*—Software effort estimation is one of the main activities in software development project, because it can indicate whether the project will be successful or failed depending on the accuracy of the estimation. There are some models to estimate effort for software development. One of the well-known models is COCOMO II. The accuracy of COCOMO II effort estimation can still be improved by optimizing its constants. Several research have been conducted to improve COCOMO II constants, such as Differential Evolution, Harmony Search and Genetic Algorithm. In this paper, a model to optimize COCOMO II constants is proposed. The optimized constants are constant A and constant B by using Bat Algorithm with Turkish project dataset. The results of experiment are compared with local calibration and COCOMO II original constants by evaluating the proposed method using Mean Magnitude Relative Error (MMRE). The results show that the proposed model has the smallest value, which means its effort estimation has closer value of actual effort than that of the other methods which exist in the Turkish project dataset.**

Keywords—software effort estimation; COCOMO II; bat algorithm; MMRE; turkish project dataset; software project

I. INTRODUCTION

Nowadays, development of software has become one of the most important thing to do for organizations. Software which has been used for long time needs to be updated and upgraded regularly to maintenance it performance. Therefore, software project estimation is gaining attention in software project management because of it importance. In the implementation, software project estimation includes several components, namely risk analysis, cost estimation, quality estimation and effort estimation [1]. When software management decisions are being made, powerful assistance of software project estimation in term of accuracy is the main key. The accuracy of effort estimation can greatly help organizations to analyze the feasibility of a project, to manage the software development process in order to reduce risks [2].

Meanwhile, inaccuracy of effort estimation can affect the software quality and the commercial reliability of software companies in a negative way [3]. On the other hand, high estimation may cause some risks for an organization, such as loss of customer and unable to compete in a competitive market place. Hence, the accuracy of effort estimation, especially in the early stages of projects needs to be improved because it is always believed to be an important goal in the software industry. For more information, the Director of OUSD, department of System Engineering and Evaluation, DR. Patricia Sanders, addressed in one of her conferences in software technology in 1998 that the organizations spent cost about $40 in software development just to rework the software. In additional, the organization went through annual loss of $18 billion in 2000. Moreover, she stated that only 16% out of 100% of software developments could be finished on time and within budget.

In fact, estimating the effort for software is difficult because of the intangible nature of the software itself. Besides that, the software usually tends to change during its development. This changes may require unexpected modification in the estimation. In last few decades, software industry has been introduced some models which can help to estimate software effort. One of the most famous software cost estimation model is Constructive Cost Model (COCOMO). COCOMO was researched and published by Barry Boehm in 1981 [4]. The model was observed from dataset which consist of 63 data points and 3 aspects of cost driver. Each data point was divided into 16 variables. Cost driver divided into Effort Multiplier (EM), Scale Factors (SF) and Line of Code (LOC) which will be used to calculate the number of effort in person-months (PM) using some formulas. Barry Boehm also proved COCOMO II with some improvements in several cost drivers in 2000 and the result was more accurate than his research before [5].

A lot of research were conducted to optimize effort estimation of COCOMO. The methods which are used in COCOMO are Differential Evolution [6], Fuzzy [7 , 8], Harmony Search [9], Water Drop [10], Genetic Algorithm [11], Tabu Search and Genetic Algorithm [12]. Meanwhile, there are algorithms, such as Bee Colony Optimization and Firefly Algorithm which are used to optimize the constants of COCOMO II [13]. In COCOMO II, there are several constants which affect the accuracy of estimation. Those constants are A, B, C and D. To get optimal constants, calibration method can be applied to local projects which has been done beforehand. Therefore, those constants can also be optimized using other methods. Metaheuristic algorithm is one of the methods which can be optimized those constants.

According to a lot of research that have been done before, the results show that the constants A and B of COCOMO II can still be improved and Metaheuristic algorithm is the algorithm which can be implemented to improve the

COCOMO II constants, because it based on population and has some equations to avoid local optimum using global search and converge to the solution using local search. In this paper, we propose a model to optimize COCOMO II constants A and B using metaheuristic search bat algorithm.

II. RELATED WORK

A lot of research have been conducted before to optimize COCOMO model. In [14], NASA dataset that contains 60 projects were processed using Harmony Search to determine the similarity measures factor of software effort and to reduce the error of MMRE. In addition, NASA software project dataset implemented the Differential Evolution method. This method was helpful to estimate the COCOMO model parameters or constants [6]. Beside that, Neural Network and Fuzzy Logic are some algorithms to optimize COCOMO model [7, 8]. The multi-layer Neural Network which works feed-forward with back-propagation are usually implemented to solve the case. Sigmoidal activation function is applied only in hidden layer while output layer uses linear function. This model was also applied in NASA dataset.

In additional, to optimize the parameters of COCOMO model, we can use Simplified Genetic Algorithm (SGA) [11]. The basic idea of SGA was also implemented on NASA software project dataset. After the SGA was implemented, the experimental results showed that this algorithm had better estimation approaching the reality over basic COCOMO. For optimizing the COCOMO II, Genetic Algorithm can also be implemented by doing post architecture model by tuning its constants to improve the accuracy of predicting both the effort and the development time. Experiments were conducted to compare the accuracy of the constants with current COCOMO II constants on NASA data set.

Intelligent Water Drops algorithm is also used to optimize the coefficients of COCOMO II model. This algorithm generates more accurate estimation in terms of effort and time for software development. This algorithm has been used in NASA 93 software project dataset [10].

Next, there are Bee Colony Optimization (BCO) algorithm and Firefly algorithm [13]. BCO algorithm were compared to Halstead model, Bailey-Basil, COCOMO II and Walston-Felix. The basic idea of BCO algorithm approach is the algorithm can generate various partial solutions and get the best solution based on Mean Magnitude of Relative Error (MMRE). This concepts have been tried in experiment and the results show that BCO algorithm is able to improve the accuracy of cost estimation and get better solution than other models. Meanwhile, Firefly algorithm is able to minimize error value significantly. Not only minimize the error value, but also show the high accuracy compared with other meta-heuristic optimization algorithms, namely Particle Swarm Optimization (PSO) [15] and Genetic Algorithms (GA). Teaching Learning Based was used in NASA 93 projects dataset to optimize constants of COCOMO II and showed increasing efficiency.

A lot of research have been conducted before to optimize COCOMO II model. Last, there is Tabu Search and Genetic Algorithm. Tabu Search algorithm, one kind of meta-heuristic algorithms and Genetic algorithm have significant role in optimization issues. This algorithm uses hybrid of Tabu Search and Genetic Algorithm to get more optimum results by using software cost estimation [12].

III. PROPOSED METHOD

In this subsection, we present the COCOMO, Bat Algorithm and our proposed method related to this research.

A. COCOMO

Boehm in 1981 introduced the research publication about the effort and schedule estimation model, called COCOMO [16]. Because of Boehm publication about COCOMO, this model becomes one of popular estimation models in 1980s.

Therefore, COCOMO in 1980's experienced difficulties to estimate the cost of software which is developed using new lifecycle processes and capabilities in new era. Therefore, there is a research to address problems on rapid development process model, non-sequential, reengineering, object oriented approaches and re-use driven approaches. Boehm introduced the new approach to COCOMO, called COCOMO II in 2000 [17]. This new approach is claimed better than COCOMO because there are some improvements in cost drivers and COCOMO II is also more accurate to solve the problems in software project estimation.

In COCOMO II, there are two models, namely Early Design and Post Architecture model. Early Design model is a high-level model in COCOMO II. This model has function to explore alternatives in architectural or do the development strategies. Meanwhile, the model which focuses on detail is called Post Architecture Model. This model works if the project is ready to be developed and maintain an existing system. The system should provide life-cycle architecture package. The detail information contains input of cost driver and cost estimates. In this research, we focus on using Post Architecture Model of COCOMO II. We implement (1) and (2).

$$PM = A \times (KLOC)^{E} \times \prod_{i=1}^{n} EM_i \quad (1)$$

where

$$E = B + 0.01 \times \sum_{j=1}^{5} SF_j \quad (2)$$

Where A and B are baseline calibration constants which was generated in calibrating in 161 projects which exist in COCOMO II dataset. Size refers to the size of the software project that is measured of Source Lines of Code in thousand (KLOC). Effort Multiplier is EM, meanwhile Scale Factor is SF.

B. Bat Algorithm

Bat algorithm is metaheuristic search that is researched and proposed by Xin-She Yang in 2010 [18]. This algorithm mimics bat behaviors using echolocation in searching prey, it can distinguish whether the object is prey, obstacle or another thing.

The need to update the bat position x_i and velocities v_i in a d-dimensional search space should be done in order to find the solution. The velocities v_i^t and new solution x_i^t at time step t are given by:

$$f_i = f_{min} + (f_{max} - f_{min})\beta$$

$$v_i^t = v_i^{t-1} + (x_i^t - x_*)f_i \quad (3)$$

$$x_i^t = x_i^{t-1} + v_i^t$$

Where:
$\beta \in [0,1]$: a random vector which obtained from uniform distribution
x_* : the current global best location after comparing all of the n bats.

The Bat Algorithm also uses local search, it make the bat can convergence to the best solution. Equation 3 shows how to generated local walk randomly.

$$x_{new} = x_{old} + \epsilon A^t \quad (4)$$

where:

$\epsilon \in [-1,1]$: random number

$A^t = \langle A_i^t \rangle$: the average loudness of all the bats at this time step.

Moreover, in every iterations process, the loudness A_i and the rate r_i of pulse emission must be updated. As the loudness usually decrease when a bat has found its prey, while the rate of pulse emission increases, we can choose the loudness as any value of convenience. For example we can use $A_{min} = 0$ and $A_0 = 1$, to indicate that a bat has just found the pray, we assuming $A_{min} = 0$ means that the bat has temporarily stop emitting any sound. The equation of the loudness A_i and the rate r_i are stated in (4).

$$A_i^{t+1} = \alpha A_i^t, \quad r_i^{t+1} = r_i^0[1 - \exp(-\gamma t)] \quad (5)$$

Where:

α and γ : constants

To move these bats towards optimal solution, their loudness and emission rates have to be updated so that the new solutions are improved.

C. Proposed Method

To optimize COCOMO II constants A and B, we use Bat algorithm to get optimal location of bats as constants A and B, by using update their location using (2), (3) and (4), then count the function by assigning it to (1) and finally count the minimum MMRE using (6).

The steps of proposed work are the following.

Input : Turkish dataset
Output : minimum MMRE of all data points

1. Start the Bat algorithm by randomizing virtual bats according to the amount of populations which have been set. The dimension of the space search is two dimension because we only search for two constants which act as bat location, namely constant A and constant B.

2. Count each estimated effort for each data points from the datasets by assigning each constants A and B which have been randomized at step 1, including Effort Multiplier and Scale Factor. Constant A and Effort Multiplier are assigned using (1) and constants B and Scale Factor are assigned using (2). The value of constants A and B are the same for 12 data points. In this step we get 12 estimated effort and we do this step for each individual bat.

3. Count the MRE for each data point by using (5). The actual effort of each data point in dataset is subtracted with estimated effort which is generated from step 2. In this step we get 12 MRE for each individual bat.

4. Count the individual evaluation MMRE using (6) by adding all 12 MRE which are produced from step 3 and divide it by 12. Run this step for all individual bats.

5. Compare the MMRE of all individual bat obtained by step 4 and find the minimum MMRE.

6. Iterate the algorithm and compare all MMRE for each iteration. The most minimum MMRE will be the optimal value of this running algorithm.

The steps of our proposed method are shown in Fig 1. The starting point is randomize virtual bats location as constants A and B, whereas the final point is obtaining the minimum MMRE.

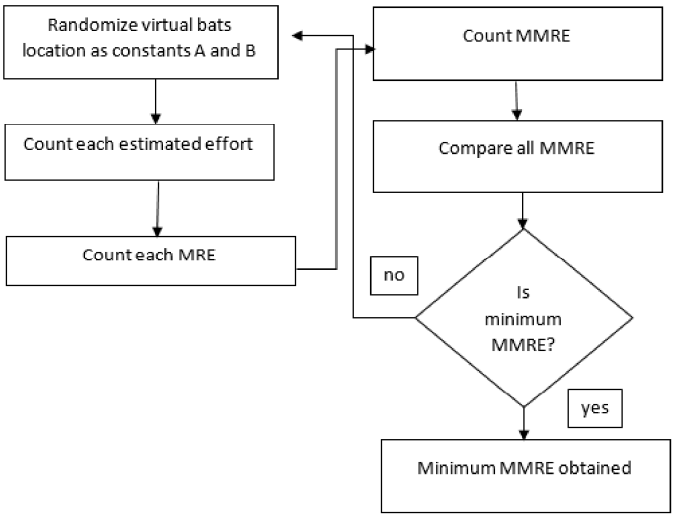


Fig. 1. The proposed method

IV. EXPERIMENT AND RESULT

A. Evaluation Criteria and Dataset

We use Magnitude Relative Error (MRE) and Mean Magnitude Relative Error (MMRE) to evaluate the result of proposed work.

$$MRE = \frac{|Actual\ Effort - Estimated\ Effort|}{Actual\ Effort} \times 100$$

$$MMRE = \frac{1}{N}\sum_{i=1}^{N} MRE_i \qquad (6)$$

The proposed work is tested using Turkish projects dataset. Some of the datasets are shown in Table I. The Turkish project dataset contains 12 projects and some attributes, namely 5 Scale Factor, 17 Effort Multiplier, Line of Code and Actual Effort.

B. Experimental Results

The aim of this experiment is to optimize constants A and B of COCOMO II estimated effort using Bat algorithm. We hypothesize according to the previous works which have been done by using meta-heuristic algorithm where using this method, we can find the optimum constants A and B. Thus, with optimum constants we will get estimated effort which more closer to actual effort by evaluate it using its MMRE. In this experiment, we use Turkish project as dataset which contains 12 projects.

The parameters setting of Bat algorithm which are used from trial and error experiment, we record the best parameters to get the best result for this experiment. The parameter settings are population = 40, iteration = 2000, loudness = 0.5, pulse rate = 0.5, minimum frequency = 0 and maximum frequency = 2.

In the experiments, we get A = 4.5644 and B = -0.23931, compared with local calibration constants, A = 3.05, and B = 0.24. We also compare with COCOMO II constants, A = 2.94, and B = 0.91. The comparison can be seen in Table II.

The results in Table II shows that MMRE which is produced by Bat algorithm optimization have the smallest value compared with MMRE of optimization using local calibration and original COCOMO II, which is 41.64% better than local calibration, and 698.89% better than COCOMO II original constants.

We also present the MRE comparison of all methods in graphic which are shown in Fig 2. In that graphic, we can see that the MRE graph of COCOMO II original constants shows fluctuated MRE compared with the other two methods, which are Bat algorithm and local calibration. MRE of Bat algorithm shows the better stability graph among the other methods.

In Fig 2, we can also get the information that dataset which has been calibrated by COCOMO II and Turkish dataset have different characteristic. The effect of using the optimized constants in project which has different characteristic can make the estimated effort not reaching or closer to its actual effort, moreover it will make the estimated effort farther away like shown in the graph.

The experiment result which is shown in Table II and Fig. 2 indicate that estimated effort by assigning constants A and B which have been optimized using Bat algorithm is closer to actual effort of projects which exist in Turkish project dataset. We can use the new constants A and B to a new project which have similar characteristic with projects in the dataset.

Optimization using Bat algorithm can get optimum result, because Bat algorithm use both local search and global search in its searching, the local search can benefit the algorithm to convergence to the optimum point, and global search can benefit the algorithm to avoid local optima.

The optimal result of Optimizing COCOMO II constants A and B using Bat Algorithm is depending on the parameters that we set. It does not always give better results, sometimes it gives bad results, because the characteristic of meta-heuristic algorithm that will find the optimal solution by searching and finding.

In the experiment of proposed work, we set the parameters by trials and errors. If we get bad result, we abandon it, otherwise we keep it and compare it with the others which have good results. The results of our proposed of each data

points does not always give the best results. We see the comparison of all the methods at data points 3, 6, 7, 9, and 12. Our proposed experiments show that the MRE results are not the best result among the other methods. This happens because the model that we proposed did not search for each optimal MRE, but we search only for optimal MMRE, which is why the MRE of each data point of our proposed work does not always give the best result among the other methods.

TABLE I. SOME OF DATA POINTS REPRESENTATION IN TURKISH SOFTWARE PROJECT DATASET

ID	1	2	3	4	5
PREC	Very High	Very High	Extra High	Extra High	Nominal
FLEX	Nominal	Nominal	Nominal	High	Nominal
RESL	Low	Low	Low	High	Low
TEAM	High	Very High	High	High	High
PMAT	Very Low	Very Low	Very Low	Nominal	Nominal
RELY	Nominal	High	High	High	Low
DATA	Nominal	Nominal	Very High	High	Nominal
CPLX	Low	Low	Low	Very High	Very High
RUSE	Low	Nominal	Nominal	Low	Nominal
DOCU	Low	Low	Low	Nominal	High
TIME	Very High	Very High	Very High	Nominal	High
STOR	High	High	Very High	Nominal	Nominal
PVOL	Nominal	Nominal	Nominal	High	Nominal
LOC	3000	2000	4250	10000	15000
AE	1.2	2	4.5	3	4

TABLE II. COMPARISON MRE AND MMRE OF BAT ALGORITHM, LOCAL CALIBRATION AND COCOMO IITABLE STYLES

Data	Bat Algorithm	Local Calibration	COCOMO II
1	27.70%	44.44%	190.67%
2	0.00%	6.87%	42.84%
3	39.13%	18.64%	106.76%
4	24.71%	151.20%	1032.58%
5	9.09%	166.88%	1478.89%
6	97.22%	89.05%	26.05%
7	64.39%	53.49%	14.43%
8	14.44%	200.28%	2841.79%
9	88.93%	28.31%	1553.36%
10	32.72%	102.47%	1499.90%
11	0.00%	22.34%	7.61%
12	12.70%	26.70%	2.75%
MMRE	**34.25%**	**75.89%**	**733.14%**

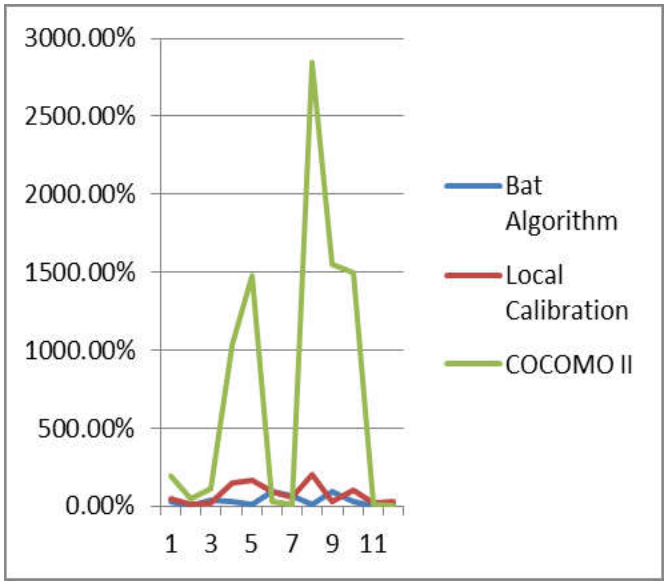


Fig. 2. Graphic MRE comparison of Bat Algorithm, Local calibration, and COCOMO II.

V. CONCLUSION

The proposed work, implementation of Bat Algorithm for COCOMO II optimization, generate better result compared with local calibration and original COCOMO II constants. We compared all models using Turkish project dataset. The experiments show that optimization using Bat algorithm can produce the smallest MMRE among the other methods. Optimization using Bat algorithm generates 41.64 % better than local calibration and 698.89 % better than COCOMO II original constants. Thus, we can assign constant A and constant B which are produced from optimization using Bat algorithm to the equation used in effort estimation. The effort estimation will produce the value which is more closer to actual effort of the projects in the dataset.

ACKNOWLEDGMENT

We would like to thank Information Management Lab, Department of Informatics, Faculty of Information and Communication Technology, Institut Teknologi Sepuluh Nopember for supporting the research.

REFERENCES

[1] M. A. Subandri and R. Sarno, "Cyclomatic Complexity for Determining Product Complexity Level in COCOMO II". *Information Systems International Conference (ISICO),* 2017. Procedia Computer Science, Volume 124, 2017, Pages 478-486.

[2] R. R. Putri, R. Sarno, D. O. Siahaan and A. Solichah. "Accuracy Improvement of The Estimations Effort in COCOMO II Based On

Logic Model of Fuzzy". *International Conference On Energy, Environment And Information System (ICENIS)*, 2016.

[3] M. Baiquni, R. Sarno, Sarwosri and Sholiq. "Improving the Accuracy of COCOMO II Using Fuzzy Logic and Local Calibration Method". *International Conference on Science in Information Technology (ICSITech),* 2017. https://doi.org/10.1109/ICSITech.2017.8257126.

[4] S. Aljahdali and A. F. Sheta, "Software Effort Estimation by Tuning COOCMO Model Parameters Using Differential Evolution,"2000.

[5] A. Malik, "An Analysis of Fuzzy Approaches for COCOMO II," no. April, pp. 68–75, 2013.

[6] K. Langsari and R. Sarno. "Optimizing COCOMO II Parameters using Particle Swarm Method". *International Conference on Science in Information Technology (ICSITech)*, 2017. https://doi.org/10.1109/ICSITech.2017.8257081.

[7] K. Langsari and R.Sarno. "Optimizing Effort and Time Parameters of COCOMO II Estimation using Fuzzy Multi-Objective PSO". *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2017. https://doi.org/10.1109/EECSI.2017.8239157.

[8] K. D. Febriyanti, R. Sarno and Y. A. Effendi, "Fraud Detection on Event Logs Using Fuzzy Association Rule Learning", *International Conference on Information & Communication Technology and Systems (ICTS),* 2017. https://doi.org/10.1109/ICTS.2017.8265661.

[9] S. Sadouni, "A New Model Of Software Cost Estimation Using Harmony Search," vol. 5, no. 3, pp. 47–55, 2015.

[10] A. G. Saif, S. Abbas, and Z. Fayed, "Intelligent Water Drops (IWD) Algorithm for COCOMO II and COQUAMO Optimization," vol.I, 2015.

[11] R. K. Sachan *et al.*, "Optimizing Basic COCOMO Model using Simplified Genetic Algorithm," *Procedia - Procedia Comput. Sci.*, vol. 89, pp. 492–498, 2016.

[12] F. Soleimanian, "A New Approach by Using Tabu Search and Genetic Algorithms in Software Cost Estimation,"2015.

[13] S. Chalotra, S. K. Sehra, Y. S. Brar, and N. Kaur, "Tuning of COCOMO Model Parameters by using Bee Colony Optimization," vol. 8, no. July, 2015.

[14] N. Ghatasheh, H. Faris, I. Aljarah, and R. M. H. Al-sayyed, "Optimizing Software Effort Estimation Models Using Firefly Algorithm," no. March, pp. 133–142, 2015.

[15] Y. A. Effendi and R. Sarno, "Discovering optimized process model using rule discovery hybrid particle swarm optimization," *3rd International Conference on Science in Information Technology (ICSITech), pp. 97-103, 2017.* DOI: 10.1109/ICSITech.2017.8257092

[16] A. Puspaningrum and R. Sarno. "Optimization of COCOMO II Coefficients using Cuckoo Optimization Algorithm to Improve The Accuracy of Effort Estimation". *International Conference on Information & Communication Technology and Systems (ICTS),* 2017. https://doi.org/10.1109/ICTS.2017.8265653.

[17] M. Alajlan, "Optimization of COCOMO II Model for Effort and Development Time Estimation using Genetic Algorithms," no. March, 2016.

[18] X. Yang, "A New Metaheuristic Bat-Inspired Algorithm," pp. 1–10, 2010.

Time and Cost Optimization Using Dynamic Programming and FMS Scheduling

Chyntia Kumalasari Puteri
Information Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya 60111, Indonesia
e-mail : chyntiaakp@gmail.com

Riyanarto Sarno
Department of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya 60111, Indonesia
e-mail : riyanarto@if.its.ac.id

Gita Intani Budiawati
Information Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya 60111, Indonesia
e-mail : gitaintani2014@gmail.com

***Abstract*—Quality of information system at a company in Surabaya which is engaged in buying and selling *wireless* devices greatly affect the business performance of the company. One of the processes that support the running of this company is the process of *Return Material Automatically (RMA)*. In this study, a RMA model business processes used by the authors for optimizing time and costs. To generate an alternative direction, the author uses an scheduling algorithm referred to Flexible Manufacturing System (FMS). After a time the optimal alternative route is obtained, it become *input* of programming objectives that will be calculated with Dynamic Programming, then the best route will be obtained using the optimal time and cost. The results of this study indicate that FMS Scheduling can be used to optimize time, and Dynamic Programming can be used to optimize the cost of this research. The results show that Trace 2 provides the optimal results compared to Trace 1.**

Keywords—Optimizing, FMS Scheduling, Dynamic Programming

I. INTRODUCTION

Along with the development of competition in the business world and the increasingly complex sales transactions as well as the broader range of sellers in the company, it has become a necessity for companies to compile, organize and discipline the jobs. The aim is to explain the details or fixed standards regarding repetitive work activities held in an organization to be able to make a better work flow..

The company that we use in this research is a company in Surabaya which is engaged in buying and selling *wireless* devices which is designated as a distributor of Ubiquiti and Mikrotik products in 2010 and a master reseller for Asia. The company itself has committed to providing quality / quality of professional products and services that ensure customer satisfaction.

Return Material Automatically (RMA) is a return transaction of damaged goods from the customer to the seller (supplier) within the warranty period / not to be repaired or replaced with other similar items / one product. RMA is a part of the process that affects the performance of companies engaged in buying and selling this wireless device. This company realizes that the time can be optimized with the right business processes in each department by accelerating the ongoing business processes and emphasizing the costs of optimizing a business process. The company's RMA has a target of achieving maximum service goods not more than 2 weeks with minimum cost.

Scheduling is needed with a view to get the high-quality time and most fee based at the referred to issues [1]. In this study, first, performed with scheduling optimization time in business processes using Flexible Manufacturing Systems (FMS). After reaching optimal time, the target to be carried out is determined [2].There are some methods used to solve the problems and give optimal result−optimization of non-linear approaches and linear optimization [3]. In the prior study, cost optimization is used using goal programming, in this study, the writer uses different method to optimize the cost, it is Dynamic Programming.

Dynamic Programming targets to optimize the cost primrily based on time, getting the quickest time, lowest cost and earnings. Goal is determined based totally on policies or SOP by the company and seeing event log / notes in that company. Dynamic Programming has been used before in a research about TSP with different constraint, and successfully used to find optimal solutions [4][5][6].

This paper is divided into several parts. Fist part explains this study problem. Second part explains about matters related to this research. The third part explains parameters and targets for FMS and Dynamic Programming methods. For the results analysis and discussion explained in fourth part. The conclusions will be explained in part five.

II. LITERATURE REVIEW

A. Scheduling

Doing a job effectively and efficiently in order to achieving the goals is what every company management desires. Scheduling is needed. Scheduling is method that aims to arrange or allocate existed resources or machine to carry out several tasks within a certain period of time [2]. Basically scheduling is a very important process to determine when an activity must begin and when the time is finished.

B. Event Log

Event log is note of all of the activity process, from the activity starts to complete. Event log basically consists of date,

time, user, and source. Every activity from a business process has special characteristic, it depends on what the organization needs to record. With event log file automatically saved in database, it helps the company to keep and resolve some problems if there is a mistake. Event log is also used for optimizing process efficiency, in several cases which reduces cost and activity time [7]. Trace from each business process has activities and each activity has its own time and cost according to the noted event log. In this study we use 2 traces; trace 1 and trace 2. The different between 2 traces is the number of activity which is done; Trace 1 has 13 activities shown in Table 1 while Trace 2 has 12 activities shown in Table 2.

TABLE I. EVENT LOG RMA DATA TRACE 1

No	Activity	Time	Cost
1	Request RMA	4	21170
2	Check Device Status	2	3670
3	Status Info	4	3670
4	Create a Service Receipt	8	22170
5	Fill Form Registration RMA	10	21170
6	Submit The Goods To The Technician	10	17500
7	Service	120	82670
8	Check Service Results via Web RMA	10	3670
9	Check Warranty	10	3670
10	Replacement	40	26170
11	Check Stock RMA	15	3670
12	Fill in the Goods Request Form	10	4870
13	Make Job Costing	10	0

TABLE II. EVENT LOG RMA DATA TRACE 2

No	Activity	Time	Cost
1	Request RMA	4	21170
2	Check Device Status	2	3670
3	Status Info	4	3670
4	Create a Service Receipt	8	22170
5	Fill Form Registration RMA	10	21170
6	Submit The Goods To The Technician	10	17500
7	Service	120	82670
8	Check Service Results via Web RMA	10	3670
9	Check Warranty	10	3670
10	Create Notice Form Damaged Goods	15	4870
11	Customer Info	15	3670
12	Make a DO RMA	30	0

C. Optimization

Optimization is parts that cannot be estranged from design process engineering. The focus of optimization is finding optimal solutions in design problem through systematic consideration of the alternatives given to resource satisfaction and cost constraints. There are many open technical problems and complex. The entire purposes of optimization are minimizing the cost, maximizing the profit, streamlining production activities, improving process efficiency, and so on. In defining the optimal solution, careful consideration of several alternatives that are often compared to several criteria is needed [8].

III. RESEARCH METHODOLOGY

A. Flexible Manufacturing Systems

FMS is job shop made automatically. Job shop is a group of machine that performs certain job from each machine center to other machine. Job shop is not only using machine but also using activity from business process. FMS is used to automatically assign alternative job shop. Several regulations in FMS are First Come First Serve (FCFS), Last Come First Serve (LCFS) or precedence [9].

Flexible Manufacturing Systems is manufacturing method producing items which might be adapted to the product adjustments being made. This system is likewise applied to calculate several cases such as distribution, maintenance, and preservation, and removal of products inside the logistics process. The entire system handles various stages of manufacturing. Flexible Manufacturing Systems is best suited for a selection of mid-variety, mid-extent manufacturing.

B. Dynamic Programming

Dynamic Programming is design techniques which is similar with divide and conquer. Dynamic programming can be applied when sub problems are not independent. Dynamic Programming algorithm solves each sub problem only once and then saves the answer in a table, thus avoiding re-computing work to obtain the answer every time the sub-problem is found [10] [11].

Dynamic programming usually applied in optimization problem. In that case, there can be several probable solutions. Every solution has value and we want to find a solution with optimal value (minimum or maximum). That solution is often referred as optimal solutions, compared with optimal solutions, because there may be several solutions that achieve optimal values [9].

On solving problems with this method: 1) There are a limited number of possible choices. 2) Solutions at each stage are built from the results of the previous stage solutions, and 3) Using optimization requirements and constraints to limit a number of choices that must be considered at a stage [13].

The steps of solving problem of *dynamic programming* are:

1. Problems can be divided into several stages, in which each stage is taken only at one decision.
2. Each stage consists of a number of states related to that stage. Generally, the state is a possible input at that stage.
3. The result of the decisions taken at each stage is transformed from the relevant status to the next status at the next stage.
4. Cost in a stage increases regularly with increasing number of stages.

5. Weight at a stage depends on the weight of the steps that have been run and the weights at that stage.
6. The best decision at a stage is independent on the decisions made in the previous stage
7. The existence of a recursive relationship that identifies the best decision for each status at stage k gives the best decision for each status in stage $k + 1$.
8. The principle of optimality applies to the problem [13].

There are two kinds of *dynamic programming*; *Forward Dynamic Programming* (calculation from front to back) and *Backward Dynamic Programming* (calculation from back to front). This study uses *Forward Dynamic Programming* method, because it begins with only 1 activity from the start. This program seeks solutions from stage 1, 2, 3, … , n. the sequence of the decision change is $x_1, x_2, x_3, \dots, x_n$. The formula of *Forward Dynamic Programming* written below [13]:

$$f_k\,(s) = c_{sxk} \quad (1)$$

$$f_k(s) = min_{x_k}\{\, c_{sxk} + f_{k+1}(x_k)\}, k = 1,2, \dots, n \quad (2)$$

Explanation:
x_k = Decision changer in stage k
c_{sxk} = Cost from s to x_k
$f_k(s, x_k)$ = Cost track total from s ke x_k
$f_k\,(s)$ = Minimum value $f_k(s, x_k)$

IV. EXPERIMENT & DISCUSSION

A. Data Collection

Data used in this study is Event log. Data collection technique in this study is event log data given directly by the company. Event Log is a process record of all activities, from start to finish taken from this company. Before being processed, this event log will be extracted and analyzed first. In this study, the log event used is 8 traces, 1 trace has 2 cases.

B. Research Flow Chart

Below is the flow chart of this study shown in Fig. 1:

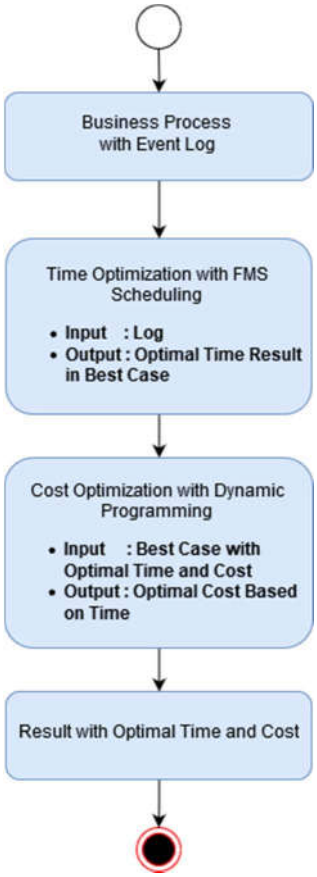


Fig. 1. Flowchart of Proposed Method

C. Time Optimization

From the given event log data, scheduling process is needed to obtain workflow of an optimal business process. Scheduling is an important process for a company. A good scheduling is one that can provide minimum time with large production in order to reduce excessive production costs. The following is a graphic showing the average time comparison of Cases and Trace 1 shown in Fig. 2 and Trace 2 shows the average time comparison of Cases and Trace 2 shown in Fig. 3 below that has been processed using *FMS Scheduling*.

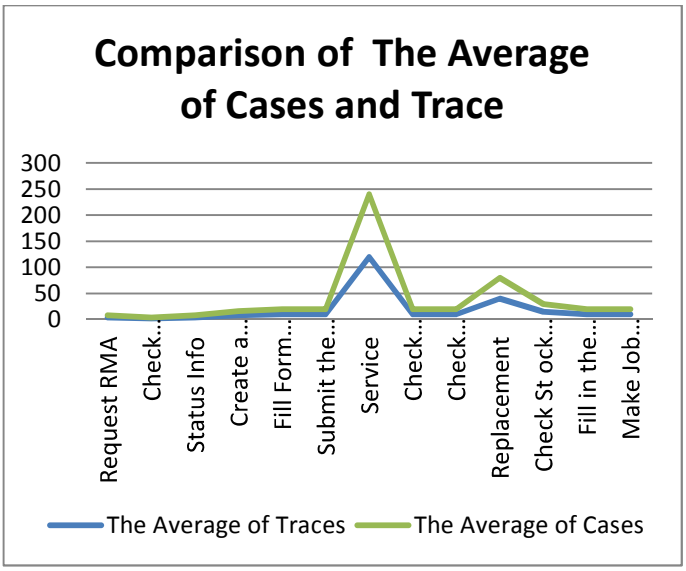


Fig. 2. Comparison of The Average of Cases and Trace 1

On Fig. above shows that Trace 1 still has not attained the desired target because there are some activities that the trace average that has not reached the average case, just as Trace 2 is shown in Fig. 3 below 2.

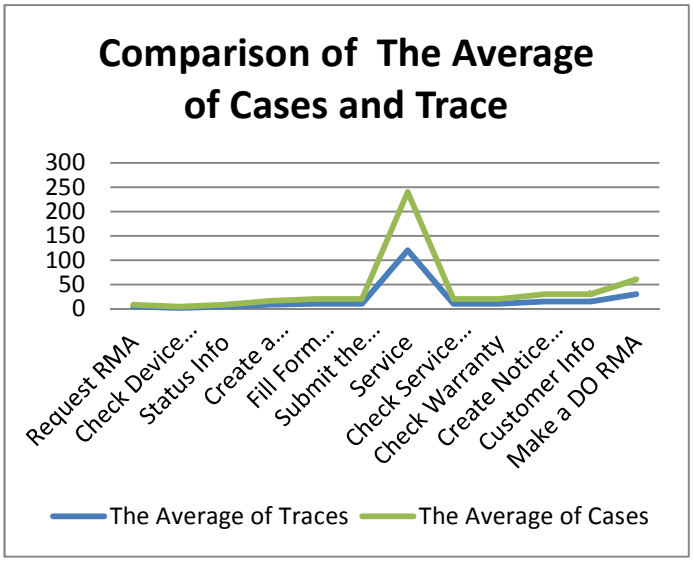


Fig. 3. Comparison of The Average of Cases and Trace 2

D. Cost Optimization

The subsequent is to optimize the cost based on optimal time done by FMS. In this stage, cost optimization is done using *Forward Dynamic Programming*. The following will show business process *Return Material Automatically (RMA)* mapped into the graphs of Trace 1 and Trace 2 that were previously processed with FMS Scheduling are shown in Fig. 4 below.

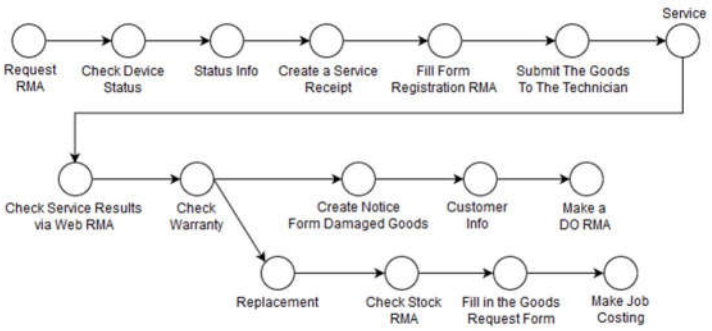


Fig. 4. Graph of Business Process Trace 1 and Trace 2

The next is making data list to calculate cost optimization by determining the value of each activity from the business process above which will be shown in Table III below.

TABLE III. LIST DATA FOR COST OPTIMIZATION

s	Activity	Cost
1	Request RMA	21170
2	Check Device Status	3670
3	Status Info	3670
4	Create a Service Receipt	22170
5	Fill Form Registration RMA	21170
6	Submit The Goods To The Technician	17500
7	Service	82670
8	Check Service Results via Web RMA	3670
9	Check Warranty	3670
10	Replacement	26170
11	Create Notice Form Damaged Goods	4870
12	Check Stock RMA	3670
13	Customer Info	3670
14	Fill in the Goods Request Form	4870
15	Make a DO RMA	0
16	Make Job Costing	0

Next is analogizing the cost of each activity from Table III above as the cost in the graph, then a more complete graph is obtained like Fig. 5 below.

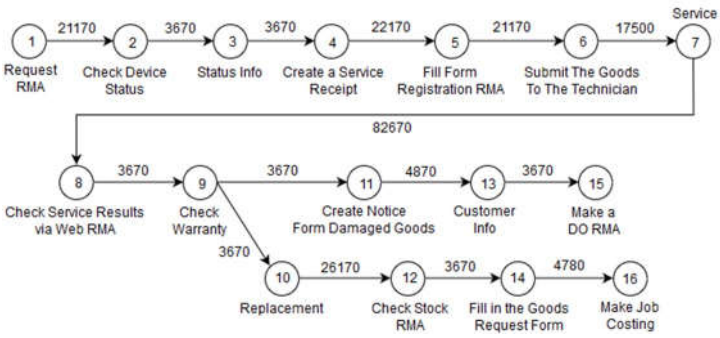


Fig. 5. Cost of Business Process Trace 1 and Trace 2

After determining the weight from each activity, calculation using *Forward Dynamic Programming* will be done with the following steps:

1) Stage 1

$$f_1(s) = c_{sx1}$$

s	$f_1(s) = c_{sx1}$	Optimum Solution	
	2	$f_1(s)$	x_1^*
1	21170	21170	2

2) Stage 2

$$f_2(s) = min_{x_2}\{ c_{sx2} + f_1(x_2)\}$$

s	$f_2(s) = c_{sx2}$	Optimum Solution	
	3	$f_2(s)$	x_2^*
2	24840	24840	3

3) Satge 3

$$f_3(s) = min_{x_3}\{ c_{sx3} + f_2(x_3)\}$$

s	$f_3(s) = c_{sx3}$	Optimum Solution	
	4	$f_3(s)$	x_3^*
3	28510	28510	4

4) Stage 4

$$f_4(s) = min_{x_4}\{ c_{sx4} + f_3(x_4)\}$$

s	$f_4(s) = c_{sx4}$	Optimum Solution	
	5	$f_4(s)$	x_4^*
4	50680	50680	5

5) Stage 5

$$f_5(s) = min_{x_5}\{ c_{sx5} + f_4(x_5)\}$$

s	$f_5(s) = c_{sx5}$	Optimum Solution	
	6	$f_5(s)$	x_5^*
5	71850	71850	6

6) Tahap6

$$f_6(s) = min_{x_6}\{ c_{sx6} + f_5(x_6)\}$$

s	$f_6(s) = c_{sx6}$	Optimum Solution	
	7	$f_6(s)$	x_6^*
6	89530	89530	7

7) Stage 7

$$f_7(s) = min_{x_7}\{ c_{sx7} + f_6(x_7)\}$$

s	$\mathbf{f_7(s) = c_{sx7}}$	Optimum Solution	
	8	$f_7(s)$	x_7^*
7	172020	172020	8

8) Stage 8

$$f_8(s) = min_{x_8}\{ c_{sx8} + f_7(x_8)\}$$

s	$\mathbf{f_8(s) = c_{sx8}}$	Optimum Solution	
	9	$f_8(s)$	x_8^*
8	175690	175690	9

9) Stage 9

$$f_9(s) = min_{x_9}\{ c_{sx9} + f_8(x_9)\}$$

s	$\mathbf{f_9(s) = c_{sx9}}$		Optimum Solution	
	10	11	$f_8(s)$	x_8^*
9	179360	179360	179360	10 or 11

10) Stage 10

$$f_{10}(s) = min_{x_{10}}\{ c_{sx10} + f_9(x_{10})\}$$

s	$\mathbf{f_{10}(s) = c_{sx10}}$		Optimum Solution	
	12	13	$f_{10}(s)$	x_{10}^*
10	205530	-	205530	13
11	-	184230	184230	

11) Stage 11

$$f_{11}(s) = min_{x_{11}}\{ c_{sx11} + f_{10}(x_{11})\}$$

s	$\mathbf{f_{11}(s) = c_{sx11}}$	Optimum Solution	
	15	$f_{11}(s)$	x_8^*
13	187900	187900	15

Cost optimization is done to find out the most optimal cost of both traces to save costs. The testing was conducted twice using similar data; the unit cost limitation was rupiah. This evaluation aims to determine the average cost, and optimal cost with the Dynamic Programming algorithm implemented.

From the process of finding the optimal cost value with Forward Dynamic Programming described above, it can be seen that the most optimum cost results from the above Business Process are Rp. 187,900 in Trace 2.

Based on the optimization time described in Fig. 2 and Fig. 3, the following results will be shown the calculation of the optimum time and the optimum cost of Trace 1 and Trace 2 shown in Table IV below.

TABLE IV. OPTIMAL TIME AND COST

Trace	Optimal Cost	Optimal Time
Trace 1	213980	60
Trace 2	187900	60

From Table IV above it is shown that it will return with a graph. For the average time result with FMS Scheduling calculation from each Trace is shown in Fig. 6 below, whereas for optimal cost result with Dynamic Programming calculation is shown in Fig. 7 below.

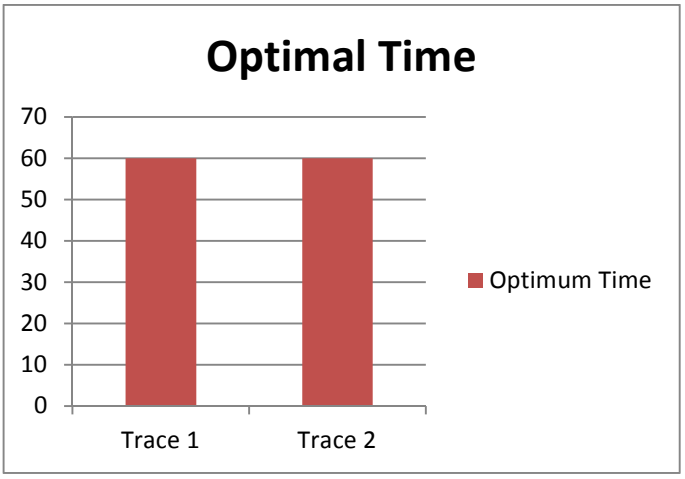


Fig. 6. Optimum Time in Each of Trace

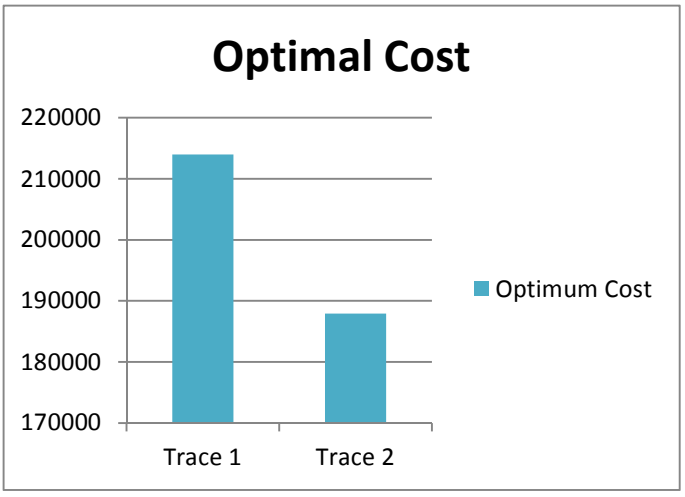


Fig. 7. Optimum Cost in Each of Trace

V. CONCLUSION

Based on the result of the calculation of time optimization with FMS Scheduling and cost optimization with Dynamic Programming on chapter IV, it can be concluded that from this study:

1. The result of time optimization on Trace 1 is 60 minutes, and the cost optimization is Rp 213,980
2. The result of time optimization on Trace 2 is 60 minutes, and the cost optimization is Rp 187,900

The results of this study can be concluded that the most optimal results is Trace 2 compared to Trace 1. Therefore, this research obtained better improvement in time optimization using FMS Scheduling and cost optimization using Dynamic Programming.

ACKNOWLEDGMENT

The authors thank to the company for providing datasets used in this research, and to the Department of Information Technology Management and the Department of Informatics, Institut Teknologi Sepuluh Nopember, Indonesia.

REFERENCES

[1] Riyanarto Sarno, Fernandes Sinaga. Business Process Anomaly Detection using Ontology-based Process Modeling and Multi-level Class Association Rule Learning. International Conference on Computer, Control, Informatics and its Applications (IC3INA). Pages: 12 – 17, 2015. DOI: 10.1109/IC3INA.2015.7377738

[2] Shoffi Izza Sabilla, Riyanarto Sarno, Yutika Amelia Effendi. Optimizing Time and Cost using Goal Programming and FMS Scheduling. 2018 International Conference on Information and Communications Technology (ICOIACT), 2018, pp. 244-249. DOI : 10.1109/ICOIACT.2018.8350727

[3] Made Agus Putra Subali, Riyanarto Sarno, Yutika Amelia Effendi. Time and Cost Optimization using Fuzzy Goal Programming. 2018 International Conference on Information and Communications Technology (ICOIACT), 2018, pp. 471-476. DOI : 10.1109/ICOIACT.2018.8350723

[4] Harilaos N Psaraftis. A dynamic programming solution to the single vehicle many-to-many immediate request dial-a-ride problem. Transportation Science, 14(2):130-154, 1980.

[5] Chryssi Malandraki and Mark S Daskin. Time dependent vehicle routing problems: For-mulations, properties and heuristic algorithms. Transportation science, 26(3):185-200,1992.

[6] A Leendert Kok, C Manuel Meyer, Herbert Kopfer, and J Marco J Schutten. A dynamic programming heuristic for the vehicle routing problem with time windows and European community social legislation. Transportation Science, 44(4):442-454, 2010.

[7] Al-Fedaghi, S., Mahdi, F. Events classification in log audit.Int J Netw Secur Appl (IJNSA), 2, 58-73, 2010.

[8] Al Haj, R. A., El-Sayegh, S. M. Timecost optimization model considering float-consumption impact. Journal of Construction Engineering and Management, 141(5), 04015001, 2015.

[9] I. Kamran, Fundamental Engineering Optimization Methods, 1sted., Kamran Iqbal & bookboon.com, 2013.

[10] Suri, R., Hildebrant, R. R. 1984. Modelling flexible manufacturingsystems using mean-value analysis. Journal of Manufacturing Systems,3(1), 27-38.

[11] Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C.. 2007. Introduction to Algorithms. Cambridge: The MIT Press.

[12] B, Paul., Agatz, Niels., Schmidt, Marie. Dynamic Programming Approaches for The Traveling Salesman Problem with Drone. Erasmus University : The Neterlands

[13] R. Gustia. 2010. Penerapan Dynamic Programming dalam MenentukanRute Penerbangan Internasional dari Indonesia. Bandung : Institut Teknologi Bandung.

Comparison Of AHP and BWM Methods Based on Geographic Information System For Determining Potential Zone of Pasir Batu Mining

Adiba Sabilla Ajrina[1]
Department of Information Technology Management, Faculty of Magister of Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
adiba.asabilla@gmail.com

Riyanarto Sarno[2]
Department of Informatics, Faculty of Information Technology and Communication
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

R.V. Hari Ginardi[3]
Department of Information Technology Management, Faculty of Magister of Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
hari@its.ac.id

***Abstract*- Infrastructure development is increasing in order to fulfill the principle of sustainable development. This has influenced the need for building materials, one of which is Pasir Batu. Kediri Regency is one of the districts that produces the largest Pasir Batu in East Java Province. This is due to the existence of several rivers which is the area of lava flow resedimentation from Mount Kelud. Currently there are already some mining activities in Kediri Regency conducted at several potential points of Pasir Batu.**

This observe aims to map the region of mining areas assessed as probably the usage of Geographic Information System (GIS) and Multi Criteria Decision Making (MCDM) using Analitycal Hierarchy Process (AHP) and Best Worst Method (BWM) methods. GIS is used for records integration in order that zoning dedication may be visual, less difficult and more powerful. Whilst MCDM is used to decide the significance and weighting criteria. The criteria used are distance from main road, distance from nearest settlement, population density, lithology condition and morphological situation.

The final result shows that the weight rank of two methods are different. In the calculation by the AHP method, we get the percentage weight as it follows : C4 (55,32%), C1 (26,72%), C5 (6,61%), C3 (6,10%) and the last is C2 (5,24%). Otherwise, in the calculation by the BWM method, we get the percentage weight as it follows : C4 (43,10%), C1 (24,60%), C2 (15,40%), C3 (12,30%) and the last is C5 (4,60%).The choice of criteria and deliberate hierarchy, additionally expert judgment to determine the level of significance of every criterion on AHP and BWM methods is very influential at the decision.

Keywords: SIG, MCDM, AHP, BWM, potential zone of mining

I. INTRODUCTION

Natural Resources are everything that comes from nature that can be used to meet the needs of human life. Based on its physical, natural resource classification consists of: (1) limited natural resources; (2) renewable natural resources; (3) Non-renewable natural resources; and (4) potential sources that are not currently a resource [1].

The natural resources of Pasir Batu in Gunung Kelud is a product of active volcanic activity. Pasir Batu from Gunung Kelud in Kediri Regency is used by the people as a mining business. Therefore, the current type of resources in this location is a renewable natural resource. In practice, there are several mining sites that are less potential after mining activities. To minimize it, it is necessary to zoning the potential of Pasir Batu mining in Kediri Regency.

No research in Indonesia has determined the potential zones of mining by using the proposed method of MCDM using AHP (Analytical Hierarchy Process) and BWM (Best Worst Method) until now. This observe aims to decide the potential zone rank of mining in Kediri Regency based on predetermined criteria. This can be used as one of the considerations by investors to determine the mining area with the desired potential. In addition, the determination of potential zones from Pasir Batu mining can be used as a basis for the government to evaluate the determination of mining areas.

The scope of this research is to determine the potential zone of mining in Kediri Regency with predetermined criteria. This study is conducted by the following study methods: spatial criteria determination, secondary data collection, analysis, calculation, and reporting.

II. PREVIOUS WORK

Over the last few years there has been a lot of research on GIS-based location selection. The research is about site selection on solar farm, wind farm, and PV site. This research is about the determination of potential zone of Pasir Batu mining that has never been done before in Indonesia.

Research on site selection is always associated with Multi Criteria Decision Making. This means it needs to be determined first about the method used. Wiguna (2016) conducted a study to optimize the selection of solar farm location using Fuzzy AHP and PROMETHEE methods. The study used two methods for weighting and ranking. This is in line with the research objectives, but at the conclusion it was revealed that the study was intended to integrate MCDM methods directly into the ArcGIS Toolbox. This integration stage is not yet clearly written on the experiment method.

The method used in this research is BWM (Best Worst Method) that is a brand new method created by means of Rezaei (2015). This method compares all criteria with the best and worst criteria. The research has been written down its calculation steps. But in the stage of determining the optimal weight is not explained how to solve it. This is quite difficult to follow the paper.

III. LITERATURE REVIEW

A. Geographic Information System

GIS is a system this is related to each other [2]. GIS as an prepared data set of pc hardware, software, geographic facts and employees designed to accumulate, save, restore, manage, examine, and show all types of environmental and geographical facts [3]. As a consequence, the analytical foundation of GIS is spatial records in virtual shape acquired through satellite data or different data digitized [4].

As previously described, Geographic Information Systems (GIS) has the ability to answer both spatial and non-spatial questions and their combinations in order to provide solutions to spatial problems [5]. This means that the system is designed to support various analyzes of geographic information, such as engineering techniques used to investigate and discover data from a spatial angle, to expand and take a look at models, and to offer data to improve knowledge [6]. It is this function or analytical technique that in Geographic Information Systems (GIS) is referred to as spatial analysis [7].

B. *Analytical Hierarchy Process (AHP)*

Analitycal Hierarchy Process (AHP) is a technique to resolve an unstructured complicated scenario into numerous components in a hierarchical arrangement, giving subjective values about the relative significance of every variable, and specifying which variable has the very best priority to persuade the final results of the scenario [8].

The selection-making technique is largely selecting the first-rate alternative. the principle system of Analitycal Hierarchy process (AHP) is to have a purposeful hierarchy with the primary enter of human belief [9]. With the hierarchy, a complicated and unstructured problem is solved into its clusters and prepared into a hierarchical shape [10].

C. *Best Worst Method (BWM)*

Best-worst method (BWM) is proposed to remedy the hassle of multi-criteria decision-making (MCDM). Some of alternatives are evaluated when it comes to some of standards for selecting the pleasant alternative [11].

In line with BWM, the best (eg maximum suitable, maximum critical) and worst (eg least suitable, least critical) criteria are recognized first through the selection maker. A pairwise assessment is then carried out among every of those criteria (best and worst) and other criteria. The maximin hassle is then formulated and solved to decide the weight of various criteria. Alternative weights with reference to distinctive criteria are received using the same technique. The very last rating of the opportunity is acquired by using combining the weights of different units of criteria and options, primarily based on which great alternative is chosen. A consistency ratio is proposed for BWM to test for assessment reliability [12].

IV. METHOD

The methodology of this research is arranged systematically as a research framework. Here are the steps of the research:

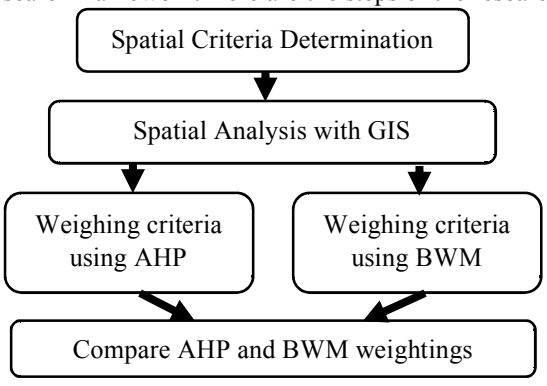


Fig. 1. Research Methodology

A. *Spatial Data Analysis*

1) Defining the Allowed Area by Regulation

a) Mining Area

Based on Keputusan Menteri ESDM No. 3672 K/30/MEM/2017 tentang Wilayah Pertambangan (WP) Pulau Jawa dan Bali, Wilayah Pertambangan (WP) is divided into Wilayah Usaha Pertambangan (WUP), Wilayah Pertambangan Rakyat (WPR) and Wilayah Pencadangan Negara (WPN). WUP is a part of WP that already has data availability, potential, and / or geological information. WPR is a part of WP where people's mining business activities are conducted. While WPN is part of WP reserved for national strategic importance.

Based on the following WP Map of Kediri Regency, almost the entire area of Kediri is WUP area, except for a small part of the easternmost area. This means that almost all areas of Kediri Regency can be proposed for mining business based on the regulation.

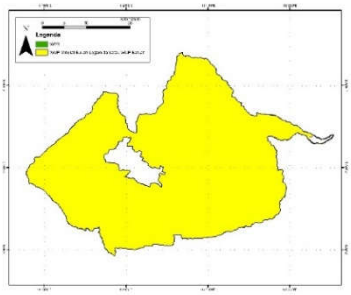


Fig. 2. Mining Area Map of Kediri Regency

b) Spatial and Region Plan

Based on Peraturan Daerah Kabupaten Kediri Nomor 14 Tahun 2011 and Peraturan Bupati Kediri Nomor 57 Tahun 2015 which regulate any area that can be mined based on Spatial and Region Plan of Kediri Regency.

Based on the Spatial and Region Plan Map of Kediri Regency, most of the permitted areas are located in the West region. The eastern region permitted is along the resedimentation area of Mount Kelud's sand which is the flow of the Ngobo and Srinjing Rivers. This is intended for mining activities to assist the government in terms of normalization of the river. However, the white areas are not permitted as a mining area.

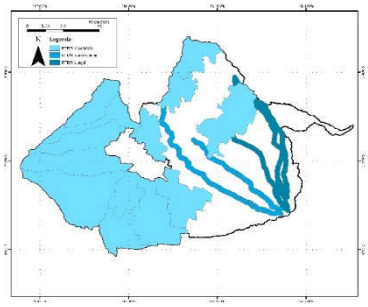


Fig. 3. Spatial and Region Plan Map of Kediri Regency

c) Forest Area

Based on Peraturan Menteri Kehutanan Republik Indonesia Nomor : P. 50/Menhut-Ii/2009 Tentang Penegasan Status Dan Fungsi Kawasan Hutan, there are 3 types of forests. They are Hutan Produksi, Hutan Lindung and Kawasan Suaka Alam. Only Hutan Produksi that can be allowed as a mining area with certain requirements. The white areas are not included in the forest map. Here is the map of Forest Area of Kediri Regency:

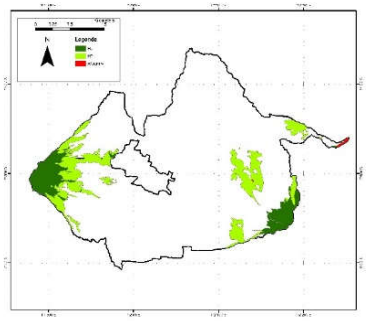


Fig. 4. Forest Area Map of Kediri Regency

2) Defining The Criteria

a) The distance from the main road

Mining activities closer to the main road will further reduce transportation costs, facilitate road access and eliminate costs to build access to mining sites. But this will also have an adverse effect on environmental and social factors. In this study, the radius area of more than 5 km from the main road got a score of 1, while the area at a radius of 2.5 km - 5 km got a score of 2. The highest score is an area with a radius of less than 2.5 km.

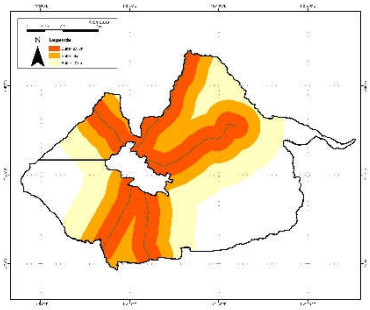


Fig. 5. The distance from the main road of Kediri Regency

b) Distance from nearby settlement

The further the mining site from the settlement the smaller the impact on the population. In this study the radius area of 1 km from the settlement got a score of 1, while the area on the radius of 1-2 km got score 2. The highest score is an area with a radius above 2 km.

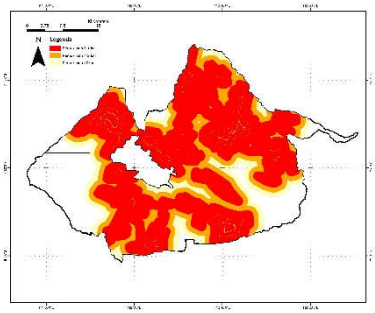


Fig. 6. The distance from nearby settlement of Kediri Regency

c) Population density

Mining activities, especially on transportation activities, will greatly affect the local environment. The smaller the population density in the mining area, the smaller the adverse impact will be. In this study, the district with more than 1,5 million people per km^2 got a score of 1, while the district with 1-1,5 million people per km^2 got a score of 2. The highest score is the district with less than 1 million people per km^2.

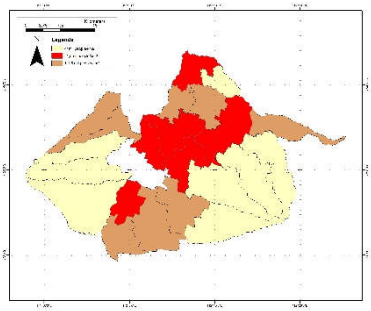


Fig. 7. Population density of Kediri Regency

d) Lithology

In Pasir Batu mining, the most suitable lithology is alluvial lithology because the material obtained is a loose material that will be easier on its mining technique. In addition, sediment lithology is also an appropriate lithology after alluvial lithology. In this study, alluvial lytology got score 3, sediment lithology got score 2 and other lithologies got score 1.

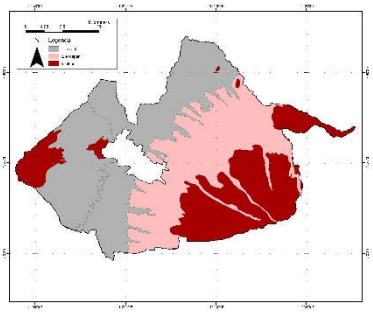


Fig. 8. Lithology of Kediri Regency

e) Morphology

The more sloping the morphological conditions at the mine site, the technical mining and security factors will be better too.This is not a major criterion but it will be easier and safer on the mining techniques performed. In this study the steep hills slopes got score 1, moderate hills slopes got score 2 and sloping hills slopes got score 3.

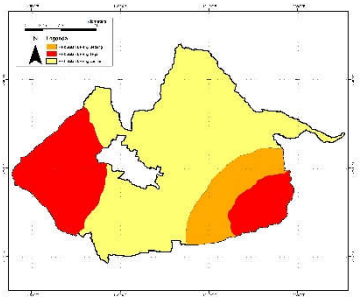


Fig. 9. Morphology of Kediri Regency

B. Determine the Weight of Each Criteria Using AHP Method

In the analysis of decision support systems using the AHP method, the first step that needs to be done is to create a hierarchical chart of each criterion. Here is a hierarchy chart on the criteria for determining the mining potential zone of Pasir Batu:

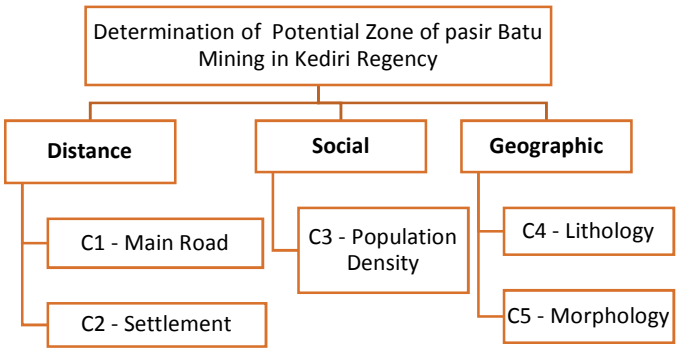


Fig. 10. Hierarchy Research Chart

To determine the priority then we must do a pairwise comparison that is comparing all the criteria for each hierarchy. The comparative assessment in this study used 9 scaled values [8]. In this study there are 5 criteria so that the results of the comparison of each criterion will produce a 5x5 matrix as in Table 1.

TABLE I. PAIRWISE COMPARISON MATRIX OF AHP

Criteria		Main Road	Settlement	Population Density	Lithology	Morphology
C1	Main Road	1,00	5,50	6,00	0,17	7,00
C2	Settlement	0,18	1,00	0,67	0,15	0,67
C3	Population Density	0,17	1,50	1,00	0,14	0,67
C4	Lithology	6,00	6,50	7,00	1,00	9,00
C5	Morphology	0,14	1,50	1,50	0,11	1,00
Amount		7,49	16,00	16,17	1,57	18,33

Based on the matrix of pairwise comparison, it is normalized by the following steps: summing the weight of each column, then the value of each column divided by the total column value, then calculating the relative priority vector of each criterion by averaging the already normalized weight. To calculate the weight of the average value (eigenvector) is by dividing a number of criteria (5). The end result of the priority weight calculation (eigenvector value) is a decimal number under one with total priorities for the criteria in one group equal to one (1). Here are the results of a normalized matrix:

TABLE II. NORMALIZED MATRIX

Criteria		Main Road	Settlement	Population Density	Lithology	Morphology	Weight
C1	Main Road	0,133	0,344	0,371	0,106	0,382	**26,72%**
C2	Settlement	0,024	0,063	0,041	0,098	0,036	**5,24%**
C3	Population Density	0,022	0,094	0,062	0,091	0,036	**6,10%**
C4	Lithology	0,801	0,406	0,433	0,635	0,491	**55,32%**
C5	Morphology	0,019	0,094	0,093	0,071	0,055	**6,61%**

Once the weights can be determined, it is necessary to calculate the consistency of the matrix. The calculation is accomplished by means of summing all of the values of every column inside the matrix, dividing every value of the column by the overall of the corresponding columns to achieve the normalization of the matrix, and summing the values of every row and dividing by the wide variety of factors to get the average. To test the consistency of the hierarchy as measured in AHP is the consistency ratio by searching on the consistency index. Consistency is predicted to be close to perfect to provide a selection that is near legitimate. Even though it is hard to obtain best, the consistency ratio is anticipated to be much less than or same to 10%. With those calculations obtained CI 0.1085 and CR 9.69% then the weighting desire is consistent.

C. Determine the Weight of Each Criteria Using BWM Method

BWM technique is used to decide the weight of every criterion. In this way, according to experts, the best criterion for determining the potential zonation of Pasir Batu mining is the condition of lithology, and the worst criterion, is the morphological condition.

TABLE III. PAIRWISE COMPARISON MATRIX OF BWM

j	Criteria	BEST Lithology a Bj	WORST Morphology a jW
j1	Main Road	2,5	6,7
j2	Settlement area	3,6	5,2
j3	Population density	6,4	2,3
j4	Lithology	1	8,5
j5	Morphology	8,5	1

Thus, the best criterion (a Bj) and the worst criterion (a jW), are compared with other criteria with a scale of 1-10. The comparison values obtained are given in Table III.

$$\min \xi$$

Subject to

$$\left|\frac{w_B}{w_1} - 2{,}5\right| \le \xi\ ,\ \left|\frac{w_B}{w_2} - 3{,}6\right| \le \xi\ ,\ ....,\ ,\ \left|\frac{w_B}{w_5} - 8{,}5\right| \le \xi$$

$$\left|\frac{w_1}{w_w} - 6{,}7\right| \le \xi\ ,\ \left|\frac{w_2}{w_w} - 5{,}2\right| \le \xi\ ,\ .....\ ,\ \left|\frac{w_4}{w_w} - 1\right| \le \xi$$

$$w_1 + w_2 + w_3 + w_4 + w_5 = 1$$

$$w_1, w_2, w_3, w_4, w_5 \ge 0 \qquad (1)$$

where:
w_{1-5}: the weight of 1-5 criterion
w_B : the weight of the best criterion
w_w : the weight of the worst criterion

The resulted model given in Eq. (1) is solved using MATLAB optimization toolbox, and weight of each criteria in Determining Potential Zone of Pasir Batu Mining were determined. Table IV shows the weights of every criterion.

TABLE IV. WEIGHT OF CRITERION USING BWM

	Criteria	Weight
j1	Main Road	24,60%
j2	Settlement area	15,40%
j3	Population density	12,30%
j4	Lithology	43,10%
j5	Morphology	4,60%

V. CONCLUSION

To support the development of increasing infrastructure, it should be supported also by appropriate resources. Resources include human and natural resources. Pasir Batu is one of the most needed commodities to support infrastructure development. Kediri Regency, as one of the sand granary of East Java Province, must know the potential points of mining area. This can be used as a basis for local governments to evaluate the determination of mining areas.

Research on location determination using MCDM method has been done but never been done to determine the potential location of Pasir Batu mining. This study uses 5 criteria as the basis for determining the potential zone of Pasir Batu mining. In the calculation by the AHP method, we get the percentage weight as it follows : C4 (55,32%), C1 (26,72%), C5 (6,61%), C3 (6,10%) and the last is C2 (5,24%). Otherwise, in the calculation by the BWM method, we get the percentage weight as it follows : C4 (43,10%), C1 (24,60%), C2 (15,40%), C3 (12,30%) and the last is C5 (4,60%). Based on the comparison of the two methods, the highest criterion is the lithology then the next criterion is the distance from the main road. In the next criteria, the two methods do not have the same sequence. The BWM Method makes the comparisons in a more structured way, which makes it easier and more understandable, and leads to more consistent comparisons, hence more reliable weights.

The selection of criteria and planned hierarchy, also expert judgment to determine the level of importance of each criterion on AHP and BWM methods is very influential on the decision. Giving weight to the priority builder is very influential on the outcome. If the range of differences between each criterion is greater then the calculation results will also be different. The level of dominance on each criterion will also differ depending on the priority value.

REFERENCES

[1] K. Wiguna, R. Sarno dan N. Ariyani, "Optimization Solar Farm Site Selection Using Multi Criteria Decision Making Fuzzy AHP and Promethee: Case Study in Bali," *International Conference on Information & Communication Technology and Systems (ICTS),* pp. 237-243, 2016.

[2] M.Sánchez-Lozano, J. Teruel-Solano, P. Soto-Elvira dan M. García-Cascales, "Geographical Information Systems (GIS) and Multi-Criteria Decision Making (MCDM) methods for the evaluation of solar farms locations: Case study in south-eastern Spain," *Renewable and Sustainable Energy Reviews,* pp. 544-556, 2013.

[3] M. Uyan, "GIS-based solar farms site selection using analytic hierarchy process," *Renewable and Sustainable Energy Reviews,* pp. 11-17, 2013.

[4] Y. Charabi dan G. A, "PV site suitability analysis using GIS-based spatial fuzzy multi-criteria evaluation," *Renewable Energy,* pp. 2554-2561, 2010.

[5] T. M., H. M. dan a. M. M., "The evaluation of solar farm locations applying Geographic Information System and Multi-Criteria Decision-Making methods: Case study in southern Morocco," *Renewable and Sustainable Energy Reviews,* pp. 1353-1362, 2015.

[6] P. Ghusti, R. Sarno dan R. H. Ginardi, "Substation Placement Optimization Method Using Delaunay Triangulation Algorithm and Voronoi Diagram In East Java Case Study," *International Conference on Information & Communication Technology and Systems (ICTS),* p. 208 – 213, 2016.

[7] J. Sánchez-Lozano, M. García-Cascales dan M. Lamata, "GIS based onshore wind farm site selection using Fuzzy Multi-Criteria Decision Making methods. Evaluating the case of Southeastern Spain," *Applied Energy,* pp. 86-102, 2016.

[8] Saaty, "How to make decision: The Analytical Hierarchy Process," *European Journal of Operational Research 48,* pp. 9-26, 1990.

[9] A. Apriliana, R. Sarno dan Y. Effendi, "Risk Analysis of IT Applications Using FMEA and AHP SAW Method With COBIT 5," *2018 International Conference on Information and Communications Technology (ICOIACT),* pp. 373 - 378, 2018.

[10] G. Kabir dan R. Sumi, "Power substation location selection using fuzzy analytic hierarchy process and PROMETHEE: A case study from Bangladesh," *Energy,* pp. 717-730, 2014.

[11] C. A. dan R. Sarno, "Application of Analytic Hierarchy Process (AHP) and Simple Additive Weighting (SAW) methods in singer selection process," *2018 International Conference on Information and Communications Technology (ICOIACT),* pp. 234-239, 2018.

[12] J. Rezaei, "Best-Worst Multi-Criteria Decision-Making Method," *Omega,* pp. 49-57, 2015.

[13] J. Rezaei, "Best-worst multi-criteria decision-making method: Some properties and a linear model," *Omega,* pp. 126-130, 2016.

Time and Cost Optimization of Parallel Services on Line Balancing Problem using Integer Programming

Tiar Anindya Putri
Department of Management Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
tiaranindyap@gmail.com

Riyanarto Sarno
Department of informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

Gita Intani Budiawati
Department of Management Technology
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
gitaintani2014@gmail.com

***Abstract*— Human resources are one of the most crucial assets of any enterprise. Optimal usage of manpower is a pivotal assignment to a manager of a dynamic control system. on this observe, we introduce a way of the concern constraints and the station limits for the parallel offerings on line Balancing problem. in lots of realistic settings, despite the fact that the real task time of a task is modeled as an growing characteristic of its starting time due to deterioration outcomes the produced man power develops constantly by means of repeated the equal or comparable activities. The technique aimed toward determining the effective size of manpower to be hired to make sure most earnings, contemplating, manufacturing or demand capacity and time of operation. This paper additionally discusses the edition of the Linear Integer Programming approach to make sure maximum earnings, deliberating, production or call for capacity and time of operation.**

Keywords—resource allocation; queue; line balancing; m/m/s; integer programming

I. INTRODUCTION

The workload to be borne in an organizational unit is closely associated with the efficiency and effectiveness of a company. that is due to the fact the workload is included inside the regulation of human resources that is an important resource within the enterprise. Many approaches that can be completed in phrases of performance of human resources, amongst others, is to further optimize the number of personnel to carry out activities appropriately [1]. Corporations sector need to strive to improve effective, efficient and flexible offerings which will innovate quick and appropriately due to the growing of the competition that results in the needs of patron desires both of best and quantity. A business process are formed from a variety of activities with various objectives . The business processes that occur in sales are related to many divisions. This can lead to waiting times, not only when the activity moves to other activities, but also additional waiting times are incurred because one worker is still doing the activity.

When a task is executed simultaneously, there is a waiting process to complete before moving on to another task. Then, when a task is executed synchronously, the next task can not be continued to another task before it finishes [3]. When the standard time of service on a business process can be optimized, then to clear up the trouble the researchers tried to implement the queue approach. Time spent on an activity to be in a queue, which are overlong, can causes the demand to be stacked up. The general purpose of this study is to determine whether the quantity of sources (server) of employee is maximized in serving queries in quality control (QC) process.

An activity in a business process calls for the assignment of resources [4]. The proper time and cost applied in order to minimize the waiting time of each activity need to be. The order in an activity becomes the main detail in the business procedure model. Moreover, the second one essential detail is the connection among those activities. There are two sorts of relationships inside the business process model, which can be parallel and sequential relationships. Sequential relationships are used to attach one activity and other activities, while parallel relations are used to attach one or extra different activities and sports. If the timestamp within the event log overlaps, then the technique version relationship is parallel [5]. Within the meantime, if the timestamp does not overlap, then the method version courting is sequential. Optimization in a business process is used to offer the highest quality, efficient and effective solution. It is able to additionally be used to clear up business process assignments [6]. this paper proposes Integer Programming for optimizing Parallel Services on Line Balancing Problem, which is a common problem in business process management. Business process optimization aims to lessen time and cost, enhance service quality, and improve client satisfaction and employees in order that the competitive benefit of an agency can be maintained.

II. LITERATURE REVIEW

A. Queue

In queuing theory, the queuing model is used to estimate the actual situation or queuing system so that queue behavior can be analyzed mathematically. The queuing model allows a number of useful steady state performance measures to be determined [7]. The achievement and queue model parameters are determined by the following notation.

λ = average arrival speed (number of arrival time unity)

$1/\lambda$ = average inter arrival time

μ = average service speed (number of units served by unity of time when server is busy)

$1/\mu$ = average time required by the server
ρ = factor of server usage (proportion of server's busy time)
Pn = the probability of n units arrivals in the system
Lq = average number of units in the queue (average queue length)
Ls = average number of units in the system
Wq = average waiting time in the queue
Ws = average waiting time in the system

Determination of the best methods to be adopted is needed to improve service delivery, reduce waiting times and optimize earnings. Arrival of goods on the business process QC company PT. XYZ occurs randomly and independently from other arrivals. Therefore, this obeys the Poisson probability distribution. The service time follows an exponential probability distribution, whilst the first come first served (FCFS) first serve queue area. The opportunity and size of model overall performance is decided analytically.

B. *Determining Probability And Performance Measures In Multi-Channel Models*

There are servers on this technique. each those servers run in parallel. They take as shown in a single channel method. control desires to do queuing studies to assist decide the satisfactory approach to lessen ready times and enhance service [2]. The look at performed in the case study at the quality control (quality controls) process confirmed that the average arrival price became 367 units consistent with hour. The probability of x arrival requests for one minute may be calculated the usage of the Poisson probability function.

$$P(x) = \frac{(\lambda^x e^{(-\lambda)})}{x!n} \quad (1)$$

Where,
x = quantity of arrivals in a given period
λ = average range of arrivals in a given duration
through the use of the exponential opportunity distribution, the chance that the service time may be less than or equal to the duration of time t can be calculated the use of the subsequent equation.

$$P(\text{services time}) \leq t = 1 - e^{(-\lambda)} \quad (2)$$

Where μ is the average of the units that can be served for a certain period

C. *Line Balancing*

Line balancing is the process of grouping work tasks in a production line into several resources by observing the balance of time and load between each work station in order to create a smooth and smooth activity process [8].

1) *Determining cycle time*
   One-cycle cycle times of QC process obtained by utilizing working hours and QC requests.
2) *Determine the minimum number of Resources and Path Efficiency*
3) *Determining mathematical model of Integer Programming for Line Balancing.*

The primary task in line balancing is to remove or lessen WIP (work in process) at bottleneck operations, maintains inventory charges low [9]. That means making better production planning by generating higher net income and allows operators to work all day to make more money and increase efficiency keep fixed prices that turn into repeat sales.

D. *Integer Programming*

Integer programming is the development of a linear program where some or all variables have integer (integer) and discrete values. Integer programming is said to be mixed or pure depending on whether some or all of these variables are limited to integer values [10]. This method is the development of linear programming method so that in the process almost have the same way. Basically, linear programming has the following count function.

1) *Linear's Function*

For the purpose of writing mathematical purposes, can be seen at the Equation (1) (2) as follows.

$$maximize = nX1 + nX2 + \cdots + nXn \quad (4)$$
$$minimize = nX1 + nX2 + \cdots + nXn \quad (5)$$

The formulation of maximize is used for purposes whose functionality improves goals and minimizes to minimize goals.
Where,
n = the positive value of the variable
$X1, X2, \ldots, Xn$ = the variable used to achieve the purpose function

2) *Constraints*

Constraints are the variables that limit a model in achieving its objectives. Writing restrictions can be seen at the Equation (6) (4) as follows.

$$Constrain\ 1 = nX1 + nX2 + \cdots + nXn \geq p \quad (6)$$
$$Constrain\ 2 = nX1 + nX2 + \cdots + nXn \geq q \quad (7)$$

Where, constraints 1 and 2 are the variables that become the boundary in achieving the objective function. While p and q are constraint values that become limiting on each boundary.

This usually applies to a corporation in figuring out the quantity of production so that you can get most profit and minimum cost [11].

The simple assembly line balancing problem is applicable for directly single product assembly traces where only precedence constraints among duties are to be taken into consideration. The simple assembly line balancing problem can be presented as the following integer programming problem.

$$\text{Min } \sum_{k=1}^{K} A_k \quad (8)$$

Subject to,

$$\sum_{k=1}^{K} t.X_{ik} \leq C, \forall k = 1,2,\ldots,K \quad (9)$$

$$\sum_{k=1}^{K} t.X_{ik} \leq C, \forall k = 1,2,\ldots,K \quad (10)$$

$$\sum_{k=1}^{K} (kx_{jk} - kx_{ik}) \geq 0, \forall i,j = 1,2,\ldots,N \text{ and } i \in P_i \quad (11)$$

$$\sum_{k=1}^{K} X_{ik} \leq nA_k, \forall k = 1,2,\ldots,K. \quad (12)$$

The objective function (8) minimizes the range of workstations. Condition (9) guarantees that every task is assigned to precisely one work station. The cycle time constraints (10) assure that the cycle time isn't always exceeded by using station time of any station. situation (11) represents the precedence constraints making sure that no task is assigned to an in advance station than a predecessor.

E. Event Log

Event logs are the procedure notes of all activities, from start to finish. Statistics or logs consist basically of date, time, user, event identification, type, and source [2]. Each activity has extraordinary attributes, relying on what the organization wishes to word. Documents are saved automatically inside the database, this enables the agency to keep and troubleshoot if there is an errors. It's also used to optimize manner performance, in a few cases lowering the fee and time of an interest [3].

III. Proposed Method

The proposed method introduced some definitions to cope with the algorithm presented in this paper. The algorithm employs this definitions together with the definitions mentioned in Literature Review. In pursuing the goal of the study, the data have been collected, classified, analyzed and the integer programming model become formulated based at the records analyzed. This QC log event will be analyzed and first using Excel Solver to see the results [12]. Because of technological and organizational conditions precedence constraints among the duties need to be located. So, the integer programming model we will put into LINGO to be optimized.

IV. Result Illustration and Analysis

As a manner to observe the conduct and to extract numerous decision parameters. The inputs are considered from the observations of a close-by offerings industry engaged in the distribution of wireless devices, as the study case. Statistics at the time consumption, mission time, precedence relationship between tasks, and the cost systems of company are acquired from the secondary data sources.

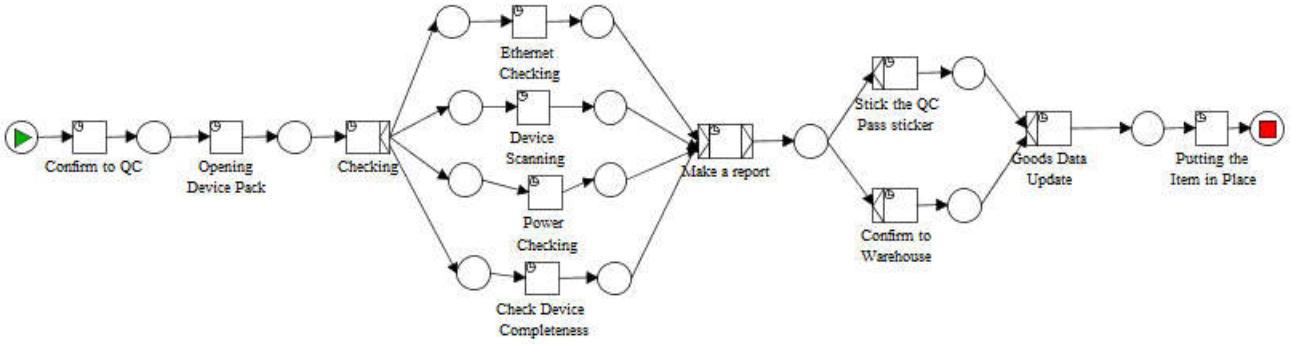


Fig 1. QC SOP in the process model by using YAWL

TABLE I. Activity Details

No	Aktivities	Cost	Duration Time	Start Time
1	Confirm to QC	143460	17.1	9:34
2	Opening Packing	3050	41.3	9:57
3	Checking	140000	41.2	10:03
4	Check Ethernet	9160	18.7	10:03
5	Scan Tool	3460	9	11:01
6	Check Power	1280	20.7	11:02
7	Check Device completeness	1700	15.2	11:01
8	Create Report	3460	70.3	12:00
9	Paste the AC Pass sticker	1720	4.5	13:15
10	Confirm to Warehouse	3460	8.3	13:15
11	Update Data Item	3460	23.9	13:40
12	Placing the Goods to the original place	3700	18.5	14:00

From the results of interviews and observations, The first step is to describe the SOP in the process model by using YAWL. It is also use for simulations, for SOP of the QC process as for Business process depiction at YAWL and setting time and cost are as follow at Figure 1.

Modeling Queue fundamentals provides the most commonplace distributions in queuing fashions, the Poisson arrival distribution and exponential service distribution. It then covers Kendall's notation, the same old technique of describing a queue model, and covers some of the fundamental queuing version assumptions. A foundation set, the equations which allow the evaluation of queue overall performance and operating characteristics are offered and associated lower back to the choice makers price minimization decision.

Table I is a breakdown of the activity, cost and duration of each activity undertaken in the QC process. Company's target of one day completion products must be finished minimal 250 stuff. Human Resources for QC team there are two people. Its time has been set on YAWL (ideal time), where everyone finished 100 pieces every day. Through computational experiments, the performance of the proposed algorithm is

examined. The experimental results validate the effectiveness and efficiency of the proposed algorithms.

TABLE II. QUEUE QUEUE TEMPLATE M/M/S

Input Data Name	Input Data	
Arrival rate	46	hour
Service rate	26	hour
Number of servers	2	for queue
Cost of Service	317910	hour
Cost of Waiting	195410	hour

Table II shows the results of initial data processing of raw data obtained. Arrival rate, service rate, number of servers provided, cost of service, cost of waiting are the variables required to perform further calculations.

After analyze optimal of, the next step is optimizing the cost based on the optimal time. As described above, the optimization of cost will use integer programming method aimed toward determining the effective size of man power to be hired to make sure most earnings.

TABLE III. TOTAL COST BASED ON QUEUE

Resource	Service Cost	Wait Cost	Total Cost
2	635820.00	1244141.01	1879961
3	953730.00	95984.87	1049715
4	1271640.00	18942.12	1290582
5	1589550.00	4062.36	1593612
6	1907460.00	840.91	1908301
7	2225370.00	162.31	2225532
8	2543280.00	28.91	2543309

The technique aimed toward determining the effective size of man power to be hired to make sure most earnings. From the calculations that have been done, the results come out with three man-power saves more cost and time savings. It saves Rp. 830.246,00 per hour of total activities. The performance characteristic of the probability (% of time) system is empty also decrease. It can be shown at the Table IV below.

TABLE IV. QUEUE PERFORMANCE CHARACTERISTICS

Queue Performance Characteristics	2 Resource	3 Resource
Average server utilization (r)	61.5% busy time	59.0% busy time
Average number of customers in the queue (L_q)	0.750 in queue	0.491 in queue
Average number of customers in the system (L)	1.981 in system	2.260 in system
Average waiting time in the queue (W_q)	0.047 hour	0.011 hour
Average time in the system (W)	0.124 hour	0.049 hour
Probability (% of time) system is empty (P_0)	0.238 empty	0.152 empty

Table IV describes the queue of performance with various assessment criteria by comparing the number of resources, namely 2 Resources and 3 Resources. From the data above, it indicates the contrast of the consequences received using line balancing queue method of solution with the present models of staff wanted. It is able to be visible in table III that using existing fashions will queue performance characteristics.

The results of this study can be used as an alternative decision recommendation to maximize the appropriate human resources, taking into account the costs and duration of work on a business process.

CONCLUSION

Based on the experimental results, the entire model that has been built can solve the problem of determining many man power with computation time is quite short. The queuing model combined with Linear Integer Programming can produce optimal results, whereby the calculations can maximize the number of employees who implement them. The completeness of the process can deliver considerable results, meeting the demands of existing working hours, and minimizing targets and limitations of the company.

ACKNOWLEDGMENT

This studies is supported by Department of Management Technology, Institut Teknologi Sepuluh Nopember. The authors are thankful to the reviewers, who have given the useful feedback and recommendations in order that our research presentation may be improved.

REFERENCES

[1] R. Sarno and S. Fernandes, "Business Process Anomaly Detection using Ontology-based Process Modeling and Multi-level Class Association Rule Learning.," *International Conference on Computer, Control, Informatics and its Applications (IC3INA).* pp. 12-17, 2015. DOI: 10.1109/IC3INA.2015.7377738

[2] J. Eze, M. Obiegbu and Jude-Eze, Statistics and Qualitative Methods for Construction and Business Managers, Nigeria: The Nigerian Institute of Building, 2005.

[3] R. Sarno, Fitrianing Haryadita, Kartini, Sarwosri, Adhatus Solichah A."Business Process Optimization from Single Timestamp Event Log," *International Conference on Computer, Control, Informatics and its Applications (IC3INA)*. Pp. 50–55, 2015. DOI: 10.1109/IC3INA.2015.7377745

[4] S. Choirunnisa, R. Sarno and A. C. Fauzan, "Optimization of Forecasted Port Container Terminal Performance Using Goal Programming," *International Conference on Information and Communications Technology (ICOIACT),* pp. 332-336, 2018. DOI: 10.1109/ICOIACT.2018.8350719

[5] R. Sarno and Y. A. Effendi, "Hierarchy Process Mining from Multi-Source Logs," *TELKOMNIKA*

Telecommunication, Computing, Electronics and Control pp. 1349-4198, 2018. DOI: http://dx.doi.org/10.12928/telkomnika.v15i4.6326

[6] S. I. Sabilla, R. Sarno and Y. A. Effendi, "Optimizing Time and Cost using Goal Programming and FMS Scheduling," *International Conference on Information and Communications Technology (ICOIACT),* pp. 244-249, 2018. DOI: 10.1109/ICOIACT.2018.8350727

[7] A. Scholl and C. Becker, "State-of-the-art exact and heuristic solution procedures for simple assembly line balancing," *European Journal of Operational Research,* vol. 168, no. 3, pp. 666-693, 2006.

[8] M. D. Toksarı, S. K. İşleyen, E. Güner and Ö. F. Baykoç, "Assembly line balancing problem with deterioration tasks and learning effect," *Expert Systems with Applications: An International Journal,* vol. 37, no. 2, pp. 1223-1228, 2010.

[9] A. Roshani and A. Roshani, "Multi-manned Assembly Line Balancing Problem: Minimizing Cycle Time," in *IIE Asian Conference*, Singapore, 2012.

[10] C. G. StallSikora and T. CantosLopesa, "Integer based formulation for the simple assembly line balancing problem with multiple identical tasks," *Computers & Operations Research,* vol. 104, no. ISSN: 0360-8352, pp. 134-144, 2017.

[11] Y. Bukchin and T. Raviv, "Constraint programming for solving various assembly line balancing problems," *OMEGA (International Journal of Management Science),* vol. 78, pp. 57-68, 2018.

[12] B. Meidyani, R. Sarno, and A.L Nurlaili, " Time and Cost Optimization Using Scheduling Job Shop and Linear Goal Programming Model," *International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 555-560. DOI: 10.1109/ICOIACT.2018.8350720.

[13] A. T. Rahman, R. Sarno, and Y. A. Effendi, "Goal Programming to Optimize Time and Cost for each Activity in Port Container Handling," *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 866-871. DOI:10.1109/ICOIACT.2018.8350808.

Integration of Fuzzy C-Means Clustering and TOPSIS (FCM-TOPSIS) with Silhouette Analysis for Multi Criteria Parameter Data

Victory Tyas P. Swindiarto
Department of Information Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
email: pambudi.17092@mhs.its.ac.id

Riyanarto Sarno
Departement of Informatics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
email: riyanarto@if.its.ac.id

Dian C. R. Novitasari
Department of Mathematics
UIN Sunan Ampel Surabaya
Surabaya, Indonesia
email: diancrini@uinsby.ac.id

***Abstract*—** Rank theory is one effective method used as an evaluation, cause in the presence of these rankings there will be competition from all aspects factor parameters. TOPSIS is excellent method in ranking. It just takes a weight and make the decision matrix. Integration required by using the best membership value in a data cluster for weighting on TOPSIS. Because the process is used for multi-criteria data, the best membership value based on the results of cluster sub-criteria to get the weight of TOPSIS. A good weight based on a valid FCM structure, silhouette coefficient needed to analyze possible displacements to other clusters. This research conducted for company performance evaluation of PT XYZ based on all branch of company.

Keywords—FCM, fuzzy, clustering, TOPSIS, silhouette, evaluation, multicriteria

I. INTRODUCTION

Ranking is one of the effective methods used as an evaluation. This rating will create competition from all aspects of factor parameters. PT XYZ is one of the companies engaged in services related to ship and transportation services. In spurring the continuity of the national industry competitiveness through the cost of competitive logistics. PT XYZ always make various cost efficiency efforts to reduce logistics costs. With competitive logistics costs, it expected that the national port industry grow and compete, both regionally and internationally. Evaluation is need to improve company performance, especially financial performance, in accordance with the expectations of the stakeholders. These performance improvement efforts are implemented using sound corporate governance principles where stakeholders play an active role as a control function. Carry out analysis and evaluation of risk exposure level of all risks across all work units within the company, based on the Key Performance Indicator (KPI) assigned in each unit.

Differences in performance appraisal of PT XYZ's branches that have several criteria. To know the value of these criteria there are several sub criteria used as indicator. The number of criteria and sub criteria in the performance appraisal in each port with 17 branches of company and head office make the performance appraisal complicated so that an approach that can accommodate all the criteria and sub criteria are needed. The approach can use the Multi Criteria Decision Making (MCDM) method. Previously PT XYZ used a comparison analysis between branches of one company to another based on one factor compared to other factors. For example, compare only in terms of income (financial) only, but in terms of infrastructure, human resources and investment values are different.

Previous Fuzzy C-Means (FCM) related studies include in FCM, data distribution observations used to perform cluster validation analysis using either K-Means or Fuzzy C-Means Method [1], the FCM Method may also combined with AHP in decision-making predictions, predicting decision impact outsourcing locations abroad on the resilience of supply chain [2]. In other fields, FCM also used to incorrectly in nuclear technology [3] and FCM analysis based on stream data to analyze misclassified [4]. FCM method used as data clustering while for ranking using TOPSIS method. The integration of the Fuzzy AHP and TOPSIS methods also provides a robust decision support system for pattern derivation suitable for cultivation and cultivation which should take precedence over priority elements [5], also used in forecasting bankruptcy of a company [6]. Used in the maritime field as a support to see the ranking there is an event at sea [7], and the most used as an evaluation / risk assessment either at the company or at school and banking [8]. Besides being used as a clustering method, it is also used for forecasting such as forecasting in maritime weather [9], it shows that fuzzy is best used in all fields. In Multi Criteria Decision Making (MCDM) is also used in the optimal selection of solar farm sites by combining two methods of fuzzy and AHP [10]. Requires a cluster validity analysis created, can use many methods such as fuzzy index [11] or by using silhouette analysis [12].

Based on the problem described there is a system for evaluation of company performance based on multi criteria using integration between Fuzzy C-Means Clustering method with TOPSIS and perform Silhouette analysis to see the strength of FCM structure. Where, a valid structure of cluster is required to analyze the presence or absence of data transfer from one cluster to another.

II. PRELIMINARIES

A. Multi Criteria Decision Making

Decision-making is based on very important principles of important factors called criteria. Criteria in a large company are not only based on one or two factors, but can be more than one factor and commonly called multi criteria. Two methods are used for multi criteria, Multi Criteria Decision Making (MCDM) and Multi Attribute Decision Making (MADM) [13].

MCDM methods include Weighted Product Method (WP), Technique for Order by Similarity to Ideal Solution Method (TOPSIS), Multi-Objective Optimization on the basis of Ratio, Profile Matching Method (PM), Simple Additive Weighting Method (SAW), Simple Multi Attribute Rating Technique Method (SMART), PROMETHEE, Analytic Hierarchy Process Method (AHP), and others.

B. Fuzzy C-Means Clustering (FCM)

FCM is a method of clustering data determined by the degree of membership. It used as clustering and classification methods. Here is the fuzzy c-means algorithm.

1. Insert data of all branches of PT XYZ based on the criteria and sub-criterion, in the form of matrix of m x n, where m is the line and n is the column.
2. Determine the number of clusters (c), Power (w), Maximum Iteration, Tolerance Error (E), Function initial objective (P0 = 0) Initial iteration (t = 1).
3. Generate a random number μ_{ik}, where: $i = 1,2,\ldots,n; k = 1,2,\ldots,c;$ as elements of the initial partition matrix called U. Then calculate the number of each column:

$$Q_i = \sum_{k=1}^{c} \mu_{ik} \quad (1)$$

Where: j = 1,2,….n

Calculate:

$$\mu_{ik} = \frac{\mu_{ik}}{Q_i} \quad (2)$$

Calculate the center of the k-cluster:

$$V_{kj} = \frac{\sum_{i=1}^{n}((\mu_{ik})^{w} * x_{ij})}{\sum_{i=1}^{n}(\mu_{ik})^{w}} \quad (3)$$

Where k = 1, 2,…, n and j = 1, 2,…,m.

4. Compute the objective function on the t-iteration,

$$P_t = \sum_{i=1}^{n}\sum_{k=1}^{c}\left(\left[\sum_{j=1}^{m}(x_{ij} - v_{kj})^2\right](\mu_{ik})^{w}\right) \quad (4)$$

5. Calculate the change in the partition matrix U by using the formula 5.

$$\mu_{ik} = \frac{\left[\sum_{j=1}^{m}(x_{ij} - v_{kj})^2\right]^{\frac{-1}{w-1}}}{\sum_{k=1}^{c}\left[\sum_{j=1}^{m}(x_{ij} - v_{kj})^2\right]^{\frac{-1}{w-1}}} \quad (5)$$

Where: i = 1, 2, .,, n and k = 1,2..,m

6. Check the stop condition:

If (Pt - Pt | <E) or (T> Maximum Iteration then stop, if not meet the condition then repeat step 4.

C. Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS)

TOPSIS uses the concept that the best alternative not only has the shortest distance from the ideal positive solution, but also the longest distance from the negative solution [14]. The TOPSIS method is one of the MCDM models used in the assessment or selection of several criteria as solution alternatives as ranking. In the TOPSIS method there is no limit on the number of attributes and alternatives used, so it can be more efficiently used to solve the problem on data that has a quantitative attribute. TOPSIS algorithm steps:

1. Create a matrix of data. Suppose the matrix (D) contains the criteria, alternatives, and weighted vectors (W) as follows.

$$D = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_m \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_n \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \end{array}$$

$$W = [w_1 \quad w_2 \quad \cdots \quad w_n]$$

Where, $C_j = C_n$ = Criteria
$A_j = A_m$ = Alternative
W = Weight Vector.

2. Make a decision matrix (n_{ij}) which is normalized by using the formula in equation 6, where i=1,2,…,m and j=1,2,…,n.

$$n_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}} \quad (6)$$

3. Make a decision matrix (y_{ij}) which is normally weighted by the equation 7 where $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$.

$$y_{ij} = w_i r_{ij} \quad (7)$$

4. Determine the positive ideal solution matrix (A ^ +) as in the following 8 equation where: j = 1,2, ..., n

$$A^{+} = (y_1^{+}, y_2^{+}, \ldots, y_n^{+}) \quad (8)$$

5. Determines the ideal solution matrix (A ^ -) as in the following equation 9 where j = 1,2, ..., n.

$$A^{-} = (y_1^{-}, y_2^{-}, \ldots, y_n^{-}) \quad (9)$$

6. Determine the distance between the value of each alternative with the positive ideal solution matrix (D_i^{+}) by the formula in the following equation 10.

$$D_i^+ = \sqrt{\sum_{j=1}^{n}(y_i^+ - y_{ij})^2} \quad (10)$$

where i = 1,2, ..., m

7. Determine the distance between the value of each alternative with the negative ideal solution matrix (D_i^-) by the formula in the following equation 11, where i = 1,2, ..., m

$$D_i^- = \sqrt{\sum_{j=1}^{n}(y_{ij} - y_i^-)^2} \quad (11)$$

where i = 1,2, ..., m

8. Determine the preference value for each alternative (V_i) by the formula in the following equation 12 where i = 1,2, ..., m.

$$V_i = \frac{D_i^-}{D_i^- + D_i^+} \quad (12)$$

D. Silhouette Coefficient

The Silhouette method used to find cluster strength or identify the quality of the cluster, and it can provide visual quality results of objects in each cluster [15]. Use information according to the number of groups in the data set (cluster). For each object denoted by the group to which it belongs, with the following 13 equations.

$$a(i) = \frac{1}{|A|-1}\sum_{j \in A, j \neq i} d(i,j) \quad (13)$$

let A be a cluster, where a (i) is the average difference of object (i) to all other objects in A. Observe the second cluster B different from A according to the formula 14.

$$d(i,B) = \frac{1}{|C|}\sum_{j \in C} d(i,j) \quad (14)$$

d (i,B) is the average difference of object (i) to all other objects in B (clusters other than cluster A), cluster B is not equal to A. After calculating d (i,B) for all B, and take the value smallest.

$$b(i) = min_{B \neq A} d(i,B) \quad (15)$$

Cluster B that reaches the minimum (d (*i,B*) = b (i)) is called the neighbor of the object (i). this is the second best cluster for the object (i). The value of Silhoutte, S(i) in object (i) is defined as in equation 16.

$$S(i) = \frac{b(i)-a(i)}{\max a(i), b(i)} \quad (16)$$

S(i): S(i): the value of the silhouette range between -1 to 1.

The value of S (i) can be explained as follows:
S (i) equal (equivalent) 1: Object (i) is classified in cluster A, and actually enters into cluster A and does not move to another cluster.
S (i) equals 0: Object (i) is located between two clusters (cluster A and cluster C), it can be said that there can be a transfer to a different cluster.
S (i) equal (equivalent) -1: object classification (i) is not exactly present in the cluster, due to overlapping. so it should be in another cluster.

III. METHODOLOGY

In this study, in general the process carried out in outline in flowchart figure 1 follows.

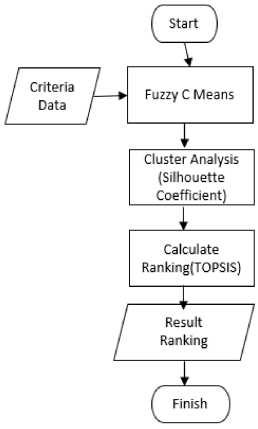


Figure 1: Flowchart of the Rank Process

Prior to the completion stage of the research, researchers identify on the research topics, namely analysis of performance evaluation of PT XYZ Branch and literature study to support the research. Then Data collection obtained at PT XYZ Head Office. The data obtained are data of company evaluation criteria based on Key Performance Indicator (KPI) that has been established in each unit and recapitulation and identify the data is needed in evaluating the performance of all branches of PT XYZ. The completion stage of this performance evaluation research is presented in Figure 1.

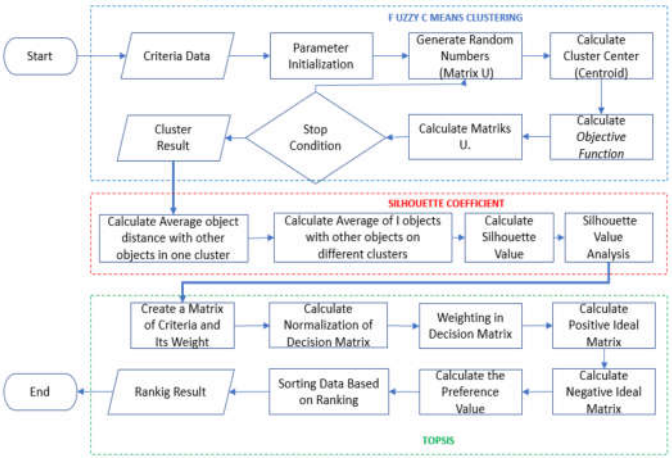


Figure 2: Flowchart Detailed Ranking Process

Beginning with data input in the form of data, clustering to get the maximum membership value of the cluster to be used as an automatic weighting on the TOPSIS process. The TOPSIS process calculates the ranking with the complete stage as in Figure 2, after calculating the normality of the decision matrix, calculate the weights in the matrix. From the result of weighted matrices, the ideal matrix is calculated for the positive solution and the ideal matrix of the negative solution and the subsequent partition matrix. Continue by calculating the preference value of each alternative. Ranking was done by sorting the value of preference from the largest to the smallest.

IV. RESULT AND ANALYSIS

PT XYZ Company has 17 Branches, there are more than 10 main criteria and more than 200 sub criteria used as Key Performance Indicator (KPI) as evaluation material of company per semester every year. In this study using five (5) main criteria with 12 sub criteria that exist. The first criterion called traffic, this traffic measures the movement of services both in and out of the port. In traffic criteria, there are five (5) sub criteria such as ship traffic, container traffic, cargo traffic, passenger traffic and animal traffic. The second criterion based on exploitation, the criteria of exploitation based on two sub criteria, business income and business operating. On the third main criteria is investment, investment is an activity company conducted to gain profit, on investment criteria based on two sub criteria that is budget and budget realization. The fourth main criterion is Cash flow, as well as investment criteria relating to the company's financial profit. The last criterion is Human Resources (HR), with two sub criteria that is budget plan and Budget Plan Realization. The entire data can seen in table 1, as Management Report PT XYZ in 2017.

To perform performance evaluation done by ranking using TOPSIS, in TOPSIS there is weighting to make decision matrix. The weight value in TOPSIS obtained from the highest value in each cluster based on the sub-parameters. Thus, the steps taken in clustering for each parameter (sub criteria) into several clusters first. To get a good cluster it needs analysis to valid or not. The method of evaluation or clustering validation used in this system is the silhouette coefficient method, to test the quality of the resulting cluster. This method is a cluster validation method that combines cohesion and separation methods. To calculate silhouette coefficient value required distance between data by using Euclidean distance formula. The first test is to see the results of silhouette on some cluster results on the parameters.

Table 1. Management Report PT XYZ in 2017

No	Branch	Traffic					Exploitation		Investment		Cash Flow	HR	
		Ship Traffic (Unit)	Container Traffic (Ton/Box)	Cargo Traffic	Passenger Traffic	Animal Traffic	Business Income	Operating Business	Budge	Realization	Profit	Budget Plane (BP)	RPB
1	A	0	196159	0	0	0	427516673000	171640244000	100026833000	124904116000	256410399000	204	185
2	B	6053	1414241	8451412	209261	9545	102505200000	80956000000	1087185000000	59041099000	258671510412	437	460
3	C	8286	202676	28576	37250	2435	193351498547	172507784510	110977000000	7391318000	21294603367	225	210
4	D	458	5459	261013	35823	3433	7389082760	6764896971	5846882539	7389082756	624225789	24	24
5	E	2392	0	2558829	96839	905	92242949000	7240544000	27018200000	2068000000	19416000000	50	49
6	F	221	0	316377	4048	224	5960000000	5810000000	9750000000	3078000000	254000000	11	11
7	G	2194	6816	6835177	26726	0	56784188084	37558101730	114885000000	5211000000	19026138824	52	52
8	H	551	20200	546635	76117	5316	21470238468	13658147267	76196000000	11121724000	1539895731	49	49
9	I	1563	50557	172803	89891	24418	37560773000	44578658000	224435000000	5850000000	-7016249	45	41
10	J	297	8795	17655	50423	0	6630289680	7189882305	7971546864	7189882305	552955606	9	9
11	K	473	11847	496578	20366	0	17627182890	16810291779	176465000000	4965000000	818529718	21	19
12	L	812	29010	473522	54891	0	26856647000	24893269000	40400000000	499510000	1963378000	46	54
13	M	779	0	11300957	0	3999	88749103000	56623303000	8846600000	36308179000	32125800000	55	55
14	N	1750	0	2230268	121008	0	96194610000	70793611000	216910000000	12878061000	23543357000	104	102
15	O	596	0	1097294	2781	293	13491623906	15272874242	93505500000	14323120000	5613973000	58	57
16	P	1793	4550	557094	407795	0	83902441498	43807196895	470308000000	5654093343	38401987549	41	43
17	Q	584	0	14664	0	0	3435783424	2835418229	696605827	545375389	545375389	17	17

Clustering test is done on the 7th parameter that is based on operating business sub criteria, tested by dividing into several clusters. The results of the experimental silhouette are determined by C (Cluster), for C = 6 in Fig. 3, C = 5 in Fig. 4, C = 4 in Fig. 5 and C = 3 is shown in Fig. 6. The results of the calculated coefficient silhouette values may vary from -1 to 1. As in Figure 3, for center cluster 6, it can be described all the values of silhouette is a positive value, there are even 4 data that is worth 1 which means the object is already in the right cluster. There is very small value in cluster 1 that is 0.0455, according to the determination Kaufman and Rousseeuw in Table 2 states that if the value is less than 0.25 it concluded no structures, which means that the data should not be on the cluster. So, the FCM structure for C = 6 is still less strong in structure. Then, more experiments on the number of other clusters should done.

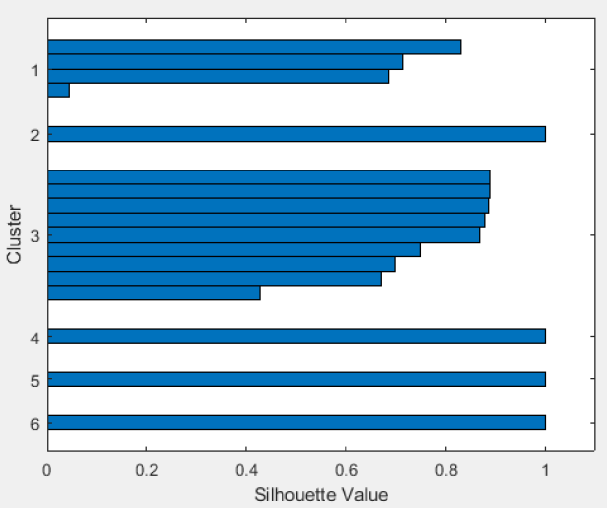


Figure 3. FCM with 6 Clusters (C = 6)

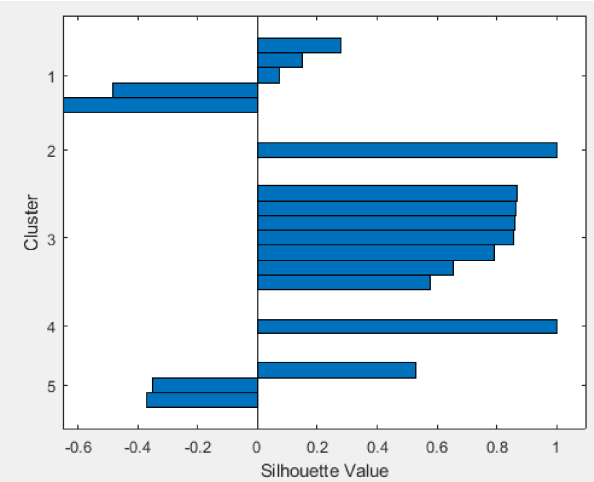


Figure 4. FCM with 5 Clusters (C = 5)

Table 2. Kaufman dan Rousseeuw Silhouette Value Table

Nilai Silhouette Coofficient	**Structure**
0.7 < SC <= 1	Strong Stucture
0.5 < SC <= 0.7	Medium Structure
0.25 < SC <= 0.5	Weak Structure
SC <= 0.25	No structure

The results of Silhouette in Figures 5 and 4 with experiments C = 5 and C = 4, suggest that the clustered data also has a much negative value of 20% of the data. From the figure, it seen that in cluster one (1) also has a relatively small silhouette value of less than 3. Which means that the data should not be included in the structure of the cluster, the silhouette value contained in other clusters in the parameter is also an average from 4 to 6, if seen in table 2, enter to the medium structure. Even in Figure 4 other than cluster 1, cluster 5 also contains a negative silhouette value. So, clusters with C = 4 and C = 5, still do not have a strong structure, even the structure is very weak.

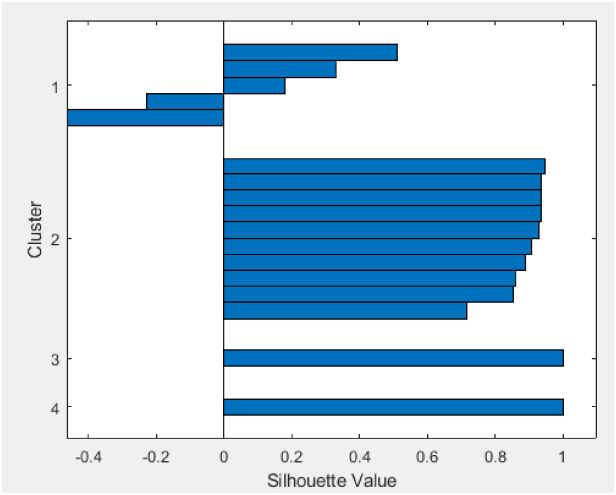


Figure 5. FCM with 4 Clusters (C = 4)

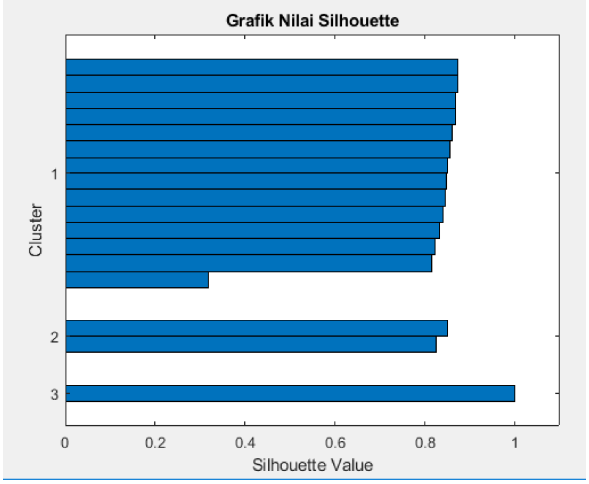


Figure 6. FCM with 3 Clusters (C = 3)

Table 3. Value of Silhoutte based on C-cluster Experiment

C=6	C=5	C=4	C=3
1.0000	1	1	0.9453
1.0000	1	1	0.6414
1.0000	0.2528	-0.2284	0.9195
0.8878	0.9054	0.9364	0.9644v
0.7494	0.7706	0.8887	0.9032
0.8859	0.9048	0.9367	0.8276
0.0455	-0.3509	0.7153	0.85
0.6713	0.7782	0.9066	0.7566
0.8304	0.5466	0.1798	0.9721
0.8878	0.9059	0.937	0.9711
0.7131	0.1678	-0.4644	0.9537
0.8679	0.8939	0.9455	0.9583
0.6980	0.7027	0.8543	0.2977
0.6861	0.6673	0.3317	0.9204
0.4267	0.6526	0.8592	0.7877
1.0000	1	0.5095	0.8827
0.8776	0.8964	0.9298	0.9566

Strong structure, seen in Fig. 6, is when FCM with C = 3, 94% of the data has a silhouette value between 0.7 and 1. According to the Kaufman and Rousseeuw, Silhouette values in Table 2, FCM cluster results have a strong structure, if it classify as a strong structure, it will be difficult possibly transfer data to other clusters. Overall, the silhouette value for each experiment performed in Table 3.

For the system, conducted the test on each parameter. Then the selected membership value on the FCM used as a weight on TOPSIS process, the results obtained as shown in Figure 7 as follows.

nilai preference	data	rangking
0.6586	2	1
0.4339	1	2
0.3519	3	3
0.2887	9	4
0.2857	16	5
0.2550	13	6
0.1857	14	7
0.1738	7	8
0.1346	5	9
0.1034	8	10
0.0706	15	11
0.0692	12	12
0.0602	11	13
0.0579	4	14
0.0405	10	15
0.0199	17	16
0.0136	6	17

Figure 7. Rank-Result FCM-TOPSIS all branch of PT XYZ

From the calculation of TOPSIS with the weighting of FCM can be analyzed the performance of the company, the ranking of the branch based on the calculation is B, A, C, I, P, M, N, G, E, H, O, L,K, D, J, Q, F and the last is F.

V. CONCLUSION

This research is an integration between clustering with ranking method, FCM and TOPSIS. Weighting by using the highest membership value on the clustering results of each criteria parameter. To analyze whether or not the clustering structure in FCM is using silhouette analysis. The result of silhouette analysis of each parameter criteria, obtained strong structure if cluster into C = 3 or C = 2. If the experiment in big C clusters, frequent overlapping or negative S (i) values. This causes there is still the possibility of moving data clustered to another cluster.

ACKNOWLEDGEMENT

Authors give a deep thank to Department of Information Technology Management Institut Teknologi Sepuluh Nopember, Department of Informatics Institut Teknologi Sepuluh Nopember and Department of Mathematics, UIN Sunan Ampel Surabaya for supporting this research.

REFERENCE

[1] K. Zhou and S. Yang, “Exploring the uniform effect of FCM clustering: A data distribution perspective,” *Knowledge-Based Syst.*, vol. 96, pp. 76–83, 2016.

[2] C. López and A. Ishizaka, “A hybrid FCM-AHP approach to predict impacts of o ff shore outsourcing location decisions on supply chain resilience,” *J. Bus. Res.*, no. February, pp. 0–1, 2017.

[3] M. Qasim Awan, L. Cao, H. Wu, W. Shen, and Z. Li, “Neutronic design study of a small modular IPWR loaded with accident tolerant-fully ceramic micro-encapsulated (AT-FCM) fuel,” *Nucl. Eng. Des.*, vol. 335, pp. 18–29, 2018.

[4] T. Gao, A. Li, and F. Meng, “Research on Data Stream Clustering Based on FCM Algorithm1,” *Procedia Comput. Sci.*, vol. 122, pp. 595–602, 2017.

[5] J. Seyedmohammadi, F. Sarmadian, A. A. Jafarzadeh, M. A. Ghorbani, and F. Shahbazi, “Application of SAW, TOPSIS and fuzzy TOPSIS models in cultivation priority planning for maize, rapeseed and soybean crops,” *Geoderma*, vol. 310, no. November 2016, pp. 178–190, 2018.

[6] D. Aloini, R. Dulmin, V. Mininno, L. Pellegrini, and G. Farina, “Technology assessment with IF-TOPSIS: An application in the advanced underwater system sector,” *Technol. Forecast. Soc. Change*, vol. 131, no. June, pp. 38–48, 2018.

[7] J. Ouenniche, B. Pérez-Gladish, and K. Bouslah, “An out-of-sample framework for TOPSIS-based classifiers with application in bankruptcy prediction,” *Technol. Forecast. Soc. Change*, vol. 131, no. April, pp. 111–116, 2018.

[8] M. Yazdi, “Risk assessment based on novel intuitionistic fuzzy-hybrid-modified TOPSIS approach,” *Saf. Sci.*, no. November 2017, pp. 0–1, 2018.

[9] D. A. Adyanti, A. H. Asyhar, D. C. R. Novitasari, A. Lubab, and M. Hafiyusholeh, “Forecasts marine weather on java sea using hybrid methods: Ts-anfis,” *Int. Conf. Electr. Eng. Comput. Sci. Informatics*, vol. 4, no. September, pp. 492–497, 2017.

[10] K. Aldrin Wiguna, R. Sarno, and N. F. Ariyani, “Optimization Solar Farm site selection using Multi-Criteria Decision Making Fuzzy AHP and PROMETHEE: case study in Bali,” *2016 Int. Conf. Inf. Commun. Technol. Syst.*, pp. 237–243, 2016.

[11] R. J. G. B. Campello and E. R. Hruschka, “A fuzzy extension of the silhouette width criterion for cluster analysis,” *Fuzzy Sets Syst.*, vol. 157, no. 21, pp. 2858–2875, 2006.

[12] N. Van Pham, L. T. Pham, T. D. Nguyen, and L. T. Ngo, “A new cluster tendency assessment method for fuzzy co-clustering in hyperspectral image analysis,” *Neurocomputing*, vol. 307, pp. 213–226, 2018.

[13] A. Cahyapratama, “Application of Analytic Hierarchy Process (AHP) and Simple Additive Weighting (SAW) Methods In Singer Selection Process,” no. Mcdm, pp. 234–239, 2018.

[14] C.-L. Hwang, Y.-J. Lai, and T.-Y. Liu, “A new approach for multiple objective decision making,” *Comput. Oper. Res.*, vol. 20, no. 8, pp. 889–899, 1993.

[15] S. R. Kannan, “A new segmentation system for brain MR images based on fuzzy techniques,” *Appl. Soft Comput. J.*, vol. 8, no. 4, pp. 1599–1606, 2008.

Performance Time Evaluation of Domestic Container Terminal using Process Mining and PERT

[1]Raka Rahardianto

Department of Management Technology, Faculty of Business and Management Technology Institut Teknologi Sepuluh November Surabaya, Indonesia

[1]raka.17092@mhs.its.ac.id

[2]Riyanarto Sarno

Department of Management Technology, Faculty of Business and Management Technology Institut Teknologi Sepuluh November Surabaya, Indonesia

[2]riyanarto@if.its.ac.id

[3]Gita Intani Budiawati

Department of Management Technology, Faculty of Business and Management Technology Institut Teknologi Sepuluh November Surabaya, Indonesia

[3]gitaintani2014@gmail.com

Abstract-Excellent Service is one of the goals of the company which is always leveled and maintained. One of the performance assessments the main service of a terminal is how fast the loading and unloading process. Evaluation of loading and unloading analyzes the number of cranes unload or load in hours, which is usually called Crane Box Hour (CBH). The performance evaluation especially the loading and unloading analyzes each process and the recorded event log in Terminal Operating System (TOS) system. Optimizing the service time is carried out by analyzing the event log using Alpha ++ and PERT Algorithms. It was shown that the problem (bottleneck) occurred clearly, and was visualized to evaluate the Critical Path.

Keywords-*Loading and Discharge Performance, Mirah Terminal, Multipurpose Terminal, Domestic, Styling, BCH (Keyword)*

I. INTRODUCTION

The terminal handles domestic container services at Tanjung Perak Port includes: PT. TPS, PT. BJTI, PT. TTL, Patchouli Terminal, and Mirah Terminal. Mirah Multipurpose Terminal main services is the container service in addition to the non-container service segment. With the spirit of change and domestic terminal services structuritation, Mirah Terminal is required to clean up and evaluate work performance of loading and unloading. With the effectiveness of loading and unloading performance, will be able to reduce logistics costs because it will reduce the cost of ship mooring, container accumulation costs in the terminal and other time-related surcharge. The basic concept of domestic container service is based on performance of ship loading and unloading services measured by Crane Box Hour (CBH)

1.1 Mirah Terminal Profile

Mirah terminal is one of the terminal under Tanjung Perak Branch management. Mirah Terminal is a multipurpose terminal serving the loading / unloading activities of Container or Non-Container (General Cargo, Bag Cargo, Dry, Motorcycle or Car). The largest business segment of this Terminal is loading and unloading services of containers [1]. Mirah Terminals have a land with a length of 640 meters, 15 meters wide, and into a 6 meter LWS pool. Performance standards set management is 10 Box or Crane / Hour, and 60 minutes Shipping Delivery Service. Mirah Terminal has a facility of 2 units of 40 Ton Tire Rubber Tyred Gantry, and Warehouse covering 13,440 M.



Fig.1 Mirah Terminal Area

1.2 Terminal Service

Mirah Terminal serves loading and unloading of containers. In the loading service process, management requested to do 75% stack first and the rest will be done by truck closing. As for unloading service done losing due to limited field capacity. Service is carried out in a package that consists of:

- **Loading Process**
  a. Loading Package
  b. Receiving Truck Loosing Package
- **Discharge Process**
  Delivery Loosing Package

In the process of handling or container handling in the port, the activities included in it are:

1. Receiving is the activity of Container movement from hinterland / out through in gate to Container yard (CY) / stacking field to stack / stacked.
2. Loading: is the Container movement activity in the Container yard (CY) / piling field to the dock to the ship to load
3. Unloading / discharge - is the activity of Container movement from ship to Container yard (CY) / stacking field to stack / stacked.

4. Stevedoring - is the work of unpacking the container from the hold of the vessel to the top of the chassis / dock or vice versa by using a ship crane or ground crane
5. Trucking / haulage - is the work of transporting a container by using a chassis within the working area of the port of the hull of the vessel to the Container yard (CY) / stacking field or vice versa.
6. Lift on - is the work of lifting the Container from the container to the top of the chassis using a transtainer / top loader or other device.
7. Lift off - is the work of lifting the Container from the top of the chasis to the stacking place by using transtainer / top loader or other tool.
8. Stacking: is the work of constructing a Container in a Container yard (CY) / stacking field or other container.

II. LITERATUR REVIEW

1.1 Standard of Loading and Unloading Performance

The Government, in this case the Ministry of Transportation of the Directorate General of Sea Communications, has issued the Regulation of the Director General of Sea Transportation number: HK.103 / 2/18 / DJPL-16 on Port Operation Performance Standards at Commercial Operated Ports. The regulation has been stipulated that Mirror Conventional Terminal BCH = 10 per Container / Crane / Hour, BSH = 10 Container / Ship / Hour, Receiving Service = 60 Minutes, and Delivery Service = 90 Minutes.

1.2 SPINER

SPINER is a Terminal Operating System called Terminal Operation System (TOS), an application used to record container transactions for domestic terminal services. The Terminal Operation System is a system where the system itself serves as a guide in performing its activities to manage all activities that run in the terminal area.

The SPINER application can record ship arrival planning, start work, request process, bill issuance process, payment process, pranota process, billing process, note issuing process integrated with ERP (ie SAP). Not only record the transactions, SPINER has Event Logs that can be derived and analyzed. It can provide data to support multi-criteria analysis decision [3]. This approach allowed for a transparent, safety, auditable and reproducible strategy in the processes of Discharge and Loading [4].

1.3 Mining Process with Alpha++ Algorithm

A method which used to mine the process sequence in the event logs is called Mining Process [5] . The algorithm used in this performance evaluation is the Alpha ++ algorithm. The Alpha ++ algorithm is an algorithm that can detect / catch an implicit relationship or dependency between processes [6]. Not only that, Alpha ++ is chosen because of its advantages compared to other Alpha algorithms. Described in the journal Van der Alast that Alpha algorithm expanded into Alpha + and expanded again to Alpha ++ [7].

1.4 PERT

Program assessment and assessment technique which is known as PERT is a network model able to mapping the completion of random activities [8]. PERT diagram has two main components of activities (activities) and milestone events / events (milestones) [9]. Both components are characterized by arcs and dots. Activities are depicted on arcs and milestones depicted on dots (circles).

The PERT making plans process consists of the subsequent steps:

- Discover precise sports (sports) and milestones,
- Decide the precise sequence of sports,
- Version a network diagram,
- Estimating the time required for each pastime,
- Determine vital tiers and pathways,
- Reveal and examine and accurate the PERT diagram for the duration of the challenge.

III. EXPERIMENT AND DISCUSSION

To know the performance of loading and unloading one cycle must be done by using:

a. Describes Business Processes using YAWL

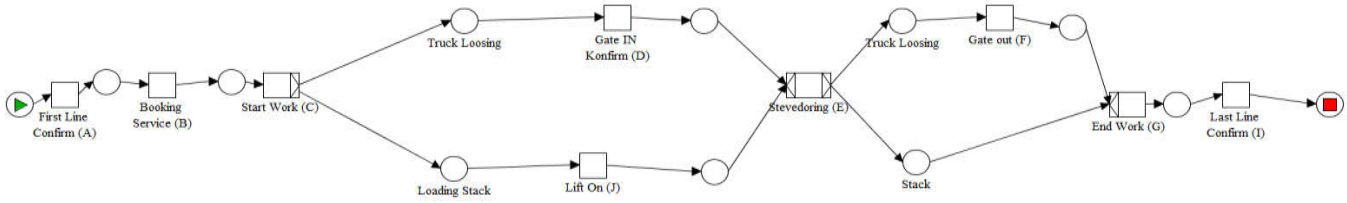


Fig.2 Domestic Terminal Business Process

The loading and unloading process of Container is generally made after confirmation from the Shipping Agent prior to the Start Working Process of Loading and Unloading [2].

Event log can be derived from Terminal Operating System (in SPINER application) [10]. The eventlog format is a text that can be converted into a csv extension file.

As follows

b. Downloading Event Log data on the system

KD_CABANG	KD_TERMINAL	CASE_ID	VOYAGE_NO	CONTAINER_NO	SERVICE_CODE	ACTIVITY	TRANS_DATE	RESOURCES	COSTS
2	2	I00000002017020200902170 9DCL020136TAKU6009628	I000000020170202009	TAKU6009628	DCL	Firstline Confirm	30/09/2017 12:00	Foreman	3500
2	2	I00000002017020200902170 9DCL020136TAKU6013741	I000000020170202009	TAKU6013741	DCL	Firstline Confirm	30/09/2017 12:00	Foreman	3500
2	2	I00000002017020200902170 9DCL020136TAKU6014408	I000000020170202009	TAKU6014408	DCL	Firstline Confirm	30/09/2017 12:00	Foreman	3500
2	2	I00000002017020200902170 9DCL020136TAKU6020545	I000000020170202009	TAKU6020545	DCL	Firstline Confirm	30/09/2017 12:00	Foreman	3500
2	2	I00000002017020200902170 9DCL020136TAKU6022471	I000000020170202009	TAKU6022471	DCL	Firstline Confirm	30/09/2017 12:00	Foreman	3500
2	2	I00000002017020200902170 9DCL020136TAKU6023713	I000000020170202009	TAKU6023713	DCL	Firstline Confirm	30/09/2017 12:00	Foreman	3500
2	2	I00000002017020200902170 9DCL020139TAKU6026040	I000000020170202009	TAKU6026040	DCL	Firstline Confirm	30/09/2017 12:00	Foreman	3500
2	2	I00000002017020200902170 9DCL020136TAKU6026190	I000000020170202009	TAKU6026190	DCL	Firstline Confirm	30/09/2017 12:00	Foreman	3500
2	2	I00000002017020200902170 9DCL020136TAKU6042077	I000000020170202009	TAKU6042077	DCL	Firstline Confirm	30/09/2017 12:00	Foreman	3500

Fig.3 Csv Extension Files from Eventlogs

From the csv file can be converted into file with. mxml.gz extension by using the help of disco application.

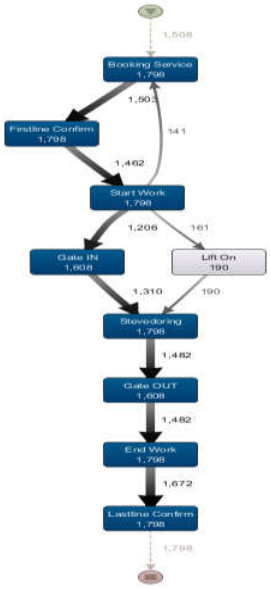


Fig.4 Result Import Data csv eventlogs.

After that, done export data form format mxml.gz. by pressing the Export Data button



Fig.5 Button Export Data



Fig.6 Export data ke format mxml.gz

Then select menu Export data into MXML (Prom 5) that this .mxml.gz file types can then be processed and carried out mining process [12].

c. Time Evaluation, using Performance Analysis with Petri Net using Alpha Algorithm.

The first process for evaluation is, that it must first be simulated business flow with Alpha ++ (Fig.7).

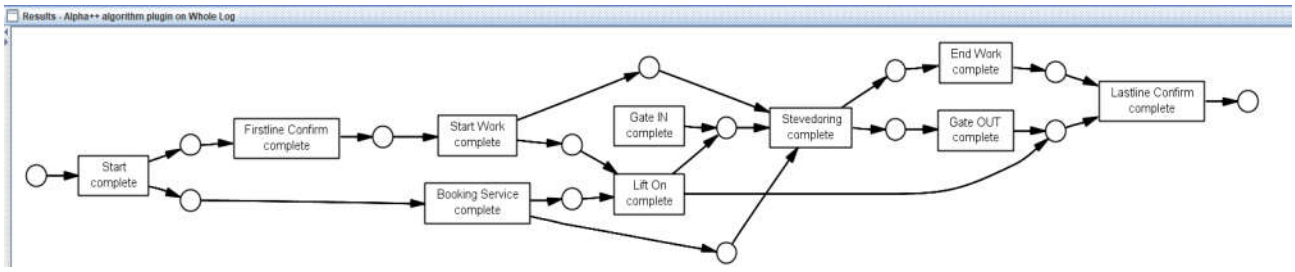


Fig.7 Business Simulation Flow with Alpha++

Thereafter, it is necessary to evaluate whether the Alpha ++ generating process is appropriate or close to the business processes of the company [11]. Furthermore, to evaluate the time in detail can be done Performance Analysis with Petri net [13]. Fig.8 is Simulation Result of Performance Analysis with Petri net [14].

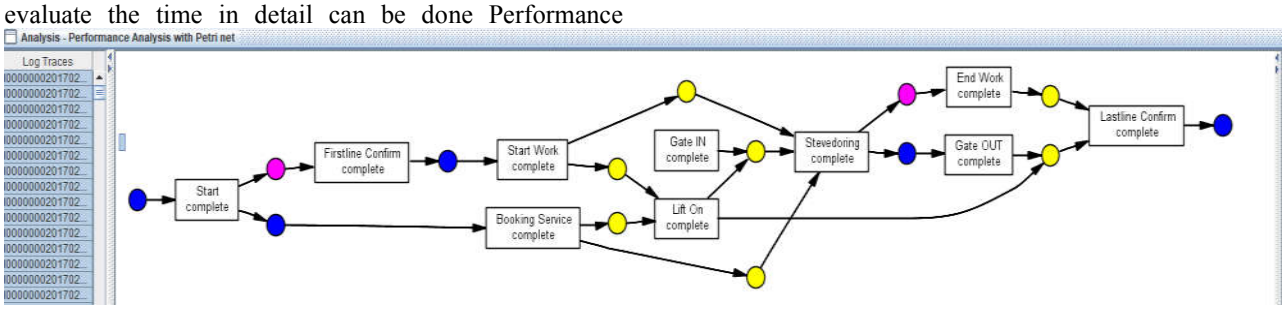


Fig.8 Simulation Performance Analysis with Petri net

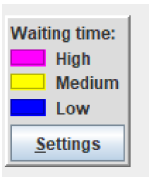


Fig.9 Waiting Time

Fig.9 this is the description of waiting Time for Simulation overall performance analysis with Petri net. With the aid of using performance assessment evaluation with Petri net, can understand the highest waiting time point [15]. There are two points that have a high waiting time that is between start complete and first line confirm complete is in Fig.10

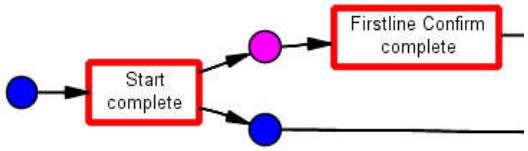


Fig.10 High Waiting Time Start Complete dan Firstline Confirm Complete.

Evaluation of transaction time between start process and first line confirm complete average 12.9 hours and the highest time is 24.75 hours. To see the detail can be seen in Fig.11 Evaluation of Transaction Time between Start process and Firstline Confirm Complete.

	Time in between (hours)
avg	12.9
min	0.0
max	24.75
stdev	9.59
fast 25.00%...	0.19
slow 25.00...	24.3
normal 50.0...	13.55

Fig.11 Evaluation of Transaction Time between Start process and Firstline Confirm Complete

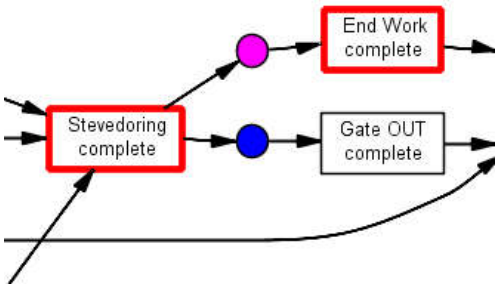


Fig.12 High Waiting Time Tinggi between Proses Stevedoring Complete process and End Work Complete

The second point observed has a high waiting time that is in Fig.12 Waiting Time High between Stevedoring Complete and End Work Complete.

Evaluation of transaction time between Stevedoring process complete and end work confirm complete average 13.93 hours and the highest time is 50.03 hours. To see the detail can be seen in Fig.13 Evaluation of Transaction Time between Stevedoring Complete and End Work Complete process.

	Time in between (hours)
avg	13.93
min	0.02
max	50.03
stdev	13.14
fast 25.00%...	2.2
slow 25.00...	33.43
normal 50.0...	10.03

Fig13 Evaluation of Transaction Time between Stevedoring Complete and End Work Complete process

d. Evaluation of Time by using Performance Sequence Diagram Analysis

Evaluation per service flow can be seen from Fig.14 which is flow of one transaction cycle. There are 15 transaction cycle patterns and have varying averages per transaction.

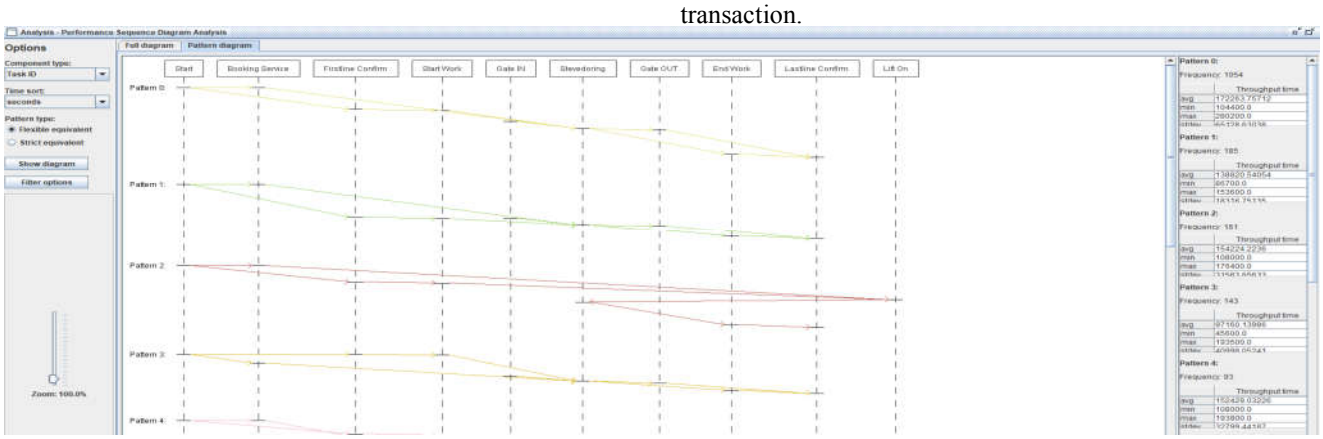


Fig.14 Average evaluation of transaction time

By using PROM it can be used to evaluate the average of service activities that is in Fig.15. Delivery Loosing service has an average of 41.99 hours, Receiving loosing has 43.31 hours on average and Loading has an average of 44.43 hours.



Fig.15 Average evaluation of transaction time per service

e. Evaluation of Pert

Calculate optimistic, general and pessimistic time of domestic container service. Fig.16 is Figure of service process flow and optimistic service time.

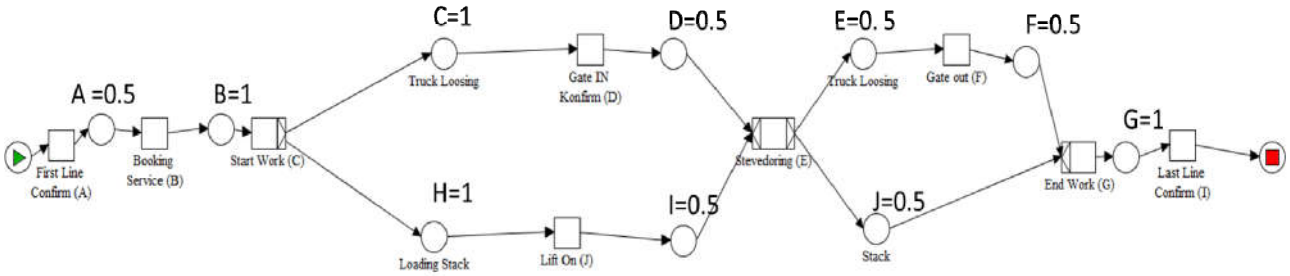


Fig.16 Service Process Flow and optimistic time

After that, the calculation is done by filling the PERT table in Fig.17. The table contains the Activity Name, previous activity, optimistic time, normal time, and pessimistic time.

Activity Number	Activity Name	Immediate Predecessor (list number/name, separated by ',')	Optimistic time (a)	Most likely time (m)	Pessimistic time (b)
1	A		0.25	0.5	1
2	B	A	0.1	1	3
3	C	B	0.3	1	2
4	D	C	0.2	0.5	1
5	E	D	0.25	0.5	2
6	F	E	0.25	0.5	1
7	G	F,J	0.3	1	2
8	H	B	0.25	1	2
9	I	H	0.1	0.5	1
10	J	I	0.3	0.5	1

Fig.17 PERT Table

From the results of the run table will show the results of evaluation as shown in Fig.17

05-13-2018 04:28:26	Activity Name	On Critical Path	Activity Mean Time	Earliest Start	Earliest Finish	Latest Start	Latest Finish	Slack (LS-ES)	Activity Time Distribution	Standard Deviation
1	A	Yes	0.5417	0	0.5417	0	0.5417	0	3-Time estimate	0.125
2	B	Yes	1.1833	0.5417	1.725	0.5417	1.725	0	3-Time estimate	0.4833
3	C	Yes	1.05	1.725	2.775	1.725	2.775	0	3-Time estimate	0.2833
4	D	Yes	0.5333	2.775	3.3083	2.775	3.3083	0	3-Time estimate	0.1333
5	E	Yes	0.7083	3.3083	4.0167	3.3083	4.0167	0	3-Time estimate	0.2917
6	F	Yes	0.5417	4.0167	4.5583	4.0167	4.5583	0	3-Time estimate	0.125
7	G	Yes	1.05	4.5583	5.6083	4.5583	5.6083	0	3-Time estimate	0.2833
8	H	no	1.0417	1.725	2.7667	2.45	3.4917	0.725	3-Time estimate	0.2917
9	I	no	0.5167	2.7667	3.2833	3.4917	4.0083	0.725	3-Time estimate	0.15
10	J	no	0.55	3.2833	3.8333	4.0083	4.5583	0.725	3-Time estimate	0.1167
	Project	Completion	Time	=	5.61	HOURs				
	Number of	Critical	Path(s)	=	1					

Fig.18 Activity Analysis PERT

PERT simulation can also be done the number of observations 1000. Fig.19 PERT simulation results can be evaluated Average service is 5.60 hours.

05-13-2018	Completion Time From	To (included)	Frequency	%	Cumulative %
0	0	1.67	0	0.0000	0.0000
1	1.67	2.06	0	0.0000	0.0000
2	2.06	2.46	0	0.0000	0.0000
3	2.46	2.85	0	0.0000	0.0000
4	2.85	3.24	0	0.0000	0.0000
5	3.24	3.64	1	0.1000	0.1000
6	3.64	4.03	4	0.4000	0.5000
7	4.03	4.43	30	3.0000	3.5000
8	4.43	4.82	89	8.9000	12.4000
9	4.82	5.21	171	17.1000	29.5000
10	5.21	5.61	225	22.5000	52.0000
11	5.61	6.00	201	20.1000	72.1000
12	6.00	6.40	148	14.8000	86.9000
13	6.40	6.79	79	7.9000	94.8000
14	6.79	7.18	38	3.8000	98.6000
15	7.18	7.58	13	1.3000	99.9000
16	7.58	7.97	1	0.1000	100.0000
17	7.97	8.37	0	0.0000	100.0000
18	8.37	8.76	0	0.0000	100.0000
19	8.76	9.15	0	0.0000	100.0000
20	9.15	9.55	0	0.0000	100.0000
21	9.55	and over	0	0.0000	100.0000
	Total	Observations =	1000	Random Seed =	27437
	Average	Completion Time =	5.60 HOURs		
	Chance	to finish in	0 HOURs	= 0.0000%	

Fig.19 PERT Results of Simulation

For the PERT evaluation method can also map, which Critical Path becomes the main time of service. Fig.20 is a Critical Path diagram PERT.

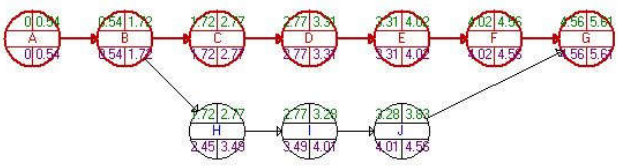


Fig.20 Critical Path PERT diagram

Fig.21 is a Figure Gantt chart with a description of the critical path of the service process.

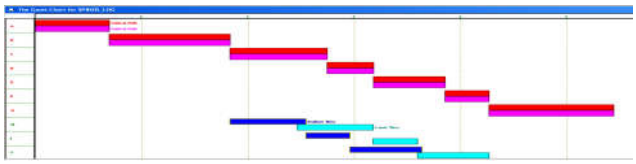
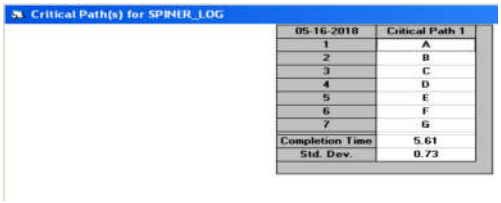


Fig.21 PERT Gantt Chart

Critical Path is a technique of analyzing network activities / activities when running a project in order to predict the total duration. A project's critical path is a sequence of activities that determines the fastest time possible for a project to be completed. With PERT we can know Critical Path that is on Fig.22 [16].



Critical Path(s) for SPINER_LOG

05-16-2018	Critical Path 1
1	A
2	B
3	C
4	D
5	E
6	F
7	G
Completion Time	5.61
Std. Dev.	0.73

Fig.22 Critical Path(s)

IV. CONCLUSIONS

From the above description and analysis, the author can make the following conclusions:

a. Evaluation by using Mining Process can be used to find out high waiting process

b. Mining process can also be evaluated at time per-service or per-activity so that it can be done evaluation or improvement.

c. Using a combination of PERT methods can be used to determine the Critical Path, which further determines which activities can be aligned or serialized (sequential)

d. For this problem, found 2 points that have a long waiting time is to start of first line and Stevedoring to end work. This can occur due to changes in ship arrival schedule, wait for the load, and the damage of loading and unloading tools.

V. REFERENCES

[1] H. M. ter Hofstede, W. M. P. van der Aalst, M. Adams, and N. Russell, Modern Business Process Automation: YAWL and its support environment. Springer Science & Business Media, 2009

[2] Riyanarto Sarno and Kelly R. Sungkono, Hidden Markov Model for Process Mining of Parallel Business Processes, International Review on Computers and Software (IRECOS), vol. 11, no. 4, pp. 290-300, 2016. DOI : 10.15866/IRECOS.V11I4.8700

[3] Nanang Yudi Setiawan, Riyanarto Sarno, Multi-Criteria Decision Making for Selecting Semantic Web Service Considering Variability and Complexity Trade-Off, Journal of Theoretical and Applied Information Technology, vol. 86, no. 2, pp. 316-326, 2016

[4] Johan Bergström, Roel van Winsen, Eder Henriqson. On the rationale of resilience in the domain of safety: A literature review. Reliability Engineering & System Safety, Volume 141, September 2015, pp 131-141. DOI : 10.1016/J.RESS.2015.03.008

[5] Riyanarto Sarno, Fernandes Sinaga. Business Process Anomaly Detection using Ontology-based Process Modeling and Multi-level Class Association Rule Learning. International Conference on Computer, Control, Informatics and its Applications (IC3INA). Pages: 12 – 17, 2015. DOI: 10.1109/IC3INA.2015.7377738

[6] Yurika Caesarita, Riyanarto Sarno, Kelly Rossa Sungkono, Identifying bottlenecks and fraud of business process using alpha ++ and heuristic miner algorithms (Case study: CV. Wicaksana Artha). 2017 11th International Conference on Information & Communication Technologyand System (ICTS), pp. 143-148. DOI: 10.1109/ICTS.2017.8265660

[7] Riyanarto Sarno, Kelly R. Sungkono, Coupled Hidden Markov Model for Process Mining of Invisible Prime Tasks, International Review on Computers and Software (IRECOS), vol. 11 no. 6, pp. 539-547, 2017. DOI : 15866/IRECOS.V11i6.9555

[8] Mete Mazlum, Ali Fuat Güneri. CPM, PERT and Project Management with Fuzzy Logic Technique and Implementation on a Business. Procedia - Social and Behavioral Sciences, Volume 210, 2 December 2015, pp 348-357. DOI: 10.1016/J.SBSPRO.2015.11.378

[9] Gita Intani Budiawati, Riyanarto Sarno, Comparison of Time Optimization of BTS Installation Project using PERT and Process Mining. International Journal of Advances in Intelligent Informatics. 2018

[10] Yurikaa Caesarita, Riyanarto Sarno, Kelly Rossa Sungkono, Identifying bottlenecks and fraud of business process using alpha ++ and heuristic miner algorithms (Case study: CV. Wicaksana Artha). International Conference on Information & Communication Technology and Systems (ICTS), 2017, pp. 143-148. DOI : https: 10.1109/ICTS.2017.8265660

[11] Riyanarto Sarno, Endang Wahyu Pamungkas, Dwi Sunaryono, Sarwosri. Business Process Composition based on Meta Models. International Seminar on Intelligent Technology and Its Applications (ISITIA),pp: 315 – 318, 2015. DOI: 10.1109/ISITIA.2015.7219998

[12] Ronny Mans, Hajo Reijers, Daniel Wismeijer, Michiel van Genuchten. Ronny Mans, Hajo Reijers, Daniel Wismeijer, Michiel van Genuchten. Information Systems, Volume 38, Issue 8, November 2013, pp 1097-1115. DOI : 10.1016/J.IS.2013.06.005

[13] Riyanarto Sarno, Bandung Arry Sanjoyo, Imam Mukhlash, Hanim Maria Astuti, Petri Net Model of ERP Business Process Variations for Small and Medium Enterprises, Journal of Theoretical and Applied Information Technology, vol. 54, no. 1, pp. 31-38, 2013

[14] Riyanarto Sarno, Bandung Arry Sanjoyo, Imam Mukhlash, Hanim Maria Astuti, Petri Net Model of ERP Business Process Variations for Small and Medium Enterprises, Journal of Theoretical and Applied Information Technology, vol. 54, no. 1, pp. 31-38, 2013

[15] Yutika Amelia Effendi, Riyanarto Sarno, Modeling Parallel Business Process Using Modified Time-based Alpha Miner. International Journal of Innovative Computing, Information and Control (IJICIC), ISSN 1349-4198

[16] Barbara Hammer, Alessio Micheli, Alessandro Sperduti, Marc Strickert, Recursive self-organizing network models. 2004 Neural Networks, Volume 17, pp 1061-1085. DOI : 10.1016/J.NEUNET.2004.06.009

An Experimental Study of Supervised Sentiment Analysis Using Gaussian Naïve Bayes

Unggul Widodo Wijayanto
Informatics Department, Faculty of Technology Information and Communication
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
unggul15@mhs.if.its.ac.id

Riyanarto Sarno
Informatics Department, Faculty of Technology Information and Communication
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
riyanarto@if.its.ac.id

***Abstract*— In this millennial generation everyone using technology at higher rates than people from other generations. It means the millennial generation is aware of evolving technology. Many companies are taking chances by receiving customer reviews through applications. In this study, we use customer reviews from Yelp (foods), IMDb (movies) and Amazon (products). The reviews received by the company are numerous. Product management does not have much time to read customer reviews one by one. So, to speed up the reading of customer reviews we were using sentiment analysis. There are many methods that used in sentiment analysis such as supervised sentiment analysis. We used TF-IDF to convert word to features implements the supervised method. Performance of the supervised method depends on the data training quality. So, to improve the accuracy of the results by improving data training quality. The methods used to improve the data training quality in this paper using CHI2 Features Selection and Stopwords. In this study, we use K-folds Cross-validation to get valid results. This study proves the use of Context-based Stopwords can improve the results. Context-based Stopwords enrich the number of Stopwords that removing bias features.**

Keywords—supervised, sentiment analysis, data training, text mining, cross-validation, features selection, stopwords.

I. INTRODUCTION

In this era, people can do anything on the internet. The Internet World Stats state that there are 4.1 billion internet users in the world on December 31, 2017 [1]. Based on internet world stats on December 31, 2017 internet user growth 1,052% for 18 years [1]. Besides using the internet to search for information they can also review. Internet users often review food, movies, and products. There are many websites that provide user reviews, like Yelp.com, IMDb.com, and Amazon.com.

Product reviews can help a company's decision-making. For the example is decision making in a cinema company. They can predict a good movie that will be of interest to many people using reviews from IMDb.com. When good reviews received massively in the first week. Cinemas can open more theaters. If the opposite happens, the cinema may reduce the duration of the theater. We can classify good and bad review by using sentiment analysis.

Some applications from Natural Language Processing (NLP) iFs opinion mining and sentiment analysis, text analysis to find out and extract the meaning of the information obtained from the review document, blogs or any source materials. And positive or negative can be the pole of a sentence [2]. Example, "This is definitely a cult classic well worth viewing and sharing with others." it means the audience has a positive sentiment to the movie. Another example, "Otherwise, don't even waste your time on this." It means the audience has a negative sentiment when watching the movie.

Sentiment analysis or usually called opinion mining had been studied many times. With the development of connectivity technology that provides many ways to interpret and process the user opinion. Some use machine learning methods such as Naïve Bayes, Maximum Entropy and Support Vector Machines [3]. In this study, we use Gaussian Naive Bayes.

This experimental study used supervised methods. Supervised methods have a good accuracy, but it depends on the data train that gives in the preprocessing. The methods used to improve the quality of data training in this paper using CHI2 features selection and Stopwords. The CHI2 algorithm is based on the χ^2 statistic to select the minimum number of attributes. Stopwords by definition are meaningless words that have low discrimination power (Lo et al., 2005) [4]. Stopwords are deleting meaningless words to avoid ambiguous results.

In this paper, we compare the correlations between the 3 datasets with the number of features selected. In addition, we also compared the effect of using General Stopwords and Context-based Stopwords. General Stopword is obtained from the function on TF-IDF. While Context-based Stopwords is obtained from each dataset.

This paper is organized as follows: Section II contains similar research has ever done; Section III explains the methods used in this study; Section IV consists of an explanation of the results; Section V consists of conclusions and future research.

II. RELATED WORKS

Research on sentiment analysis by classifying customer reviews using machine learning or lexicon-based has been widely practiced.

S. Rani and P. Kumar [5] conducting research on sentiment analysis system which aims to improve teaching and learning. This research uses natural language processing and machine learning to analyze student feedbacks.

Y. M. Aye and S. S. Aung [6] research sentiment analysis customer review on a restaurant in Myanmar in the form of text by using Lexicon-based.

III. Approach

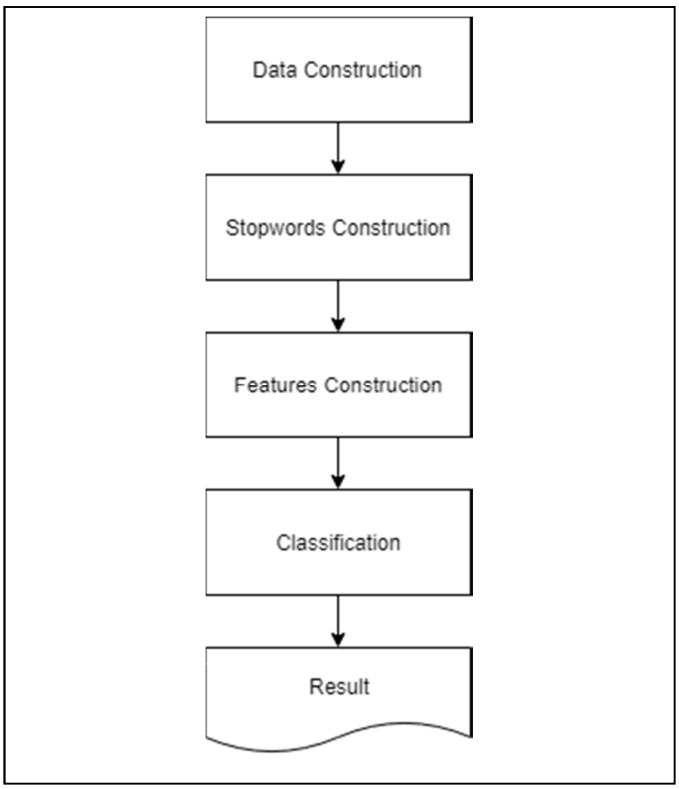


Fig. 1. Research Overview Diagram

Fig. 1 shows the flow from data construction to sentiment classification. The flow is divided by 4 part, the first part data construction, the second part is Stopwords Construction, and the third part features construction divided by 2 different processes that we want to compare using TF-IDF Stopwords and using Context-based Stopwords. The results of each algorithm used will be sorted and compared based on which has the highest scores on the F-score and accuracy.

A. Data Construction



Fig. 2. Data Construction Diagram

There are 3 datasets used in this study. The dataset is a customer review of 3 websites that have different topics. The first topic is the food reviews obtained from the yelp.com website. The second topic is film reviews obtained from IMDb.com website. The third topic is the product reviews obtained from the amazon.com website. 3 datasets are obtained from The UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences) [9]. The dataset consists of customer reviews and labels sentiments. Each dataset has been assigned a positive (1) or negative (0) label based on the sentiments of each review. Each dataset consists of 500 positives and 500 negatives. Fig. 2 shows how the dataset is formed.

B. Stopwords Construction

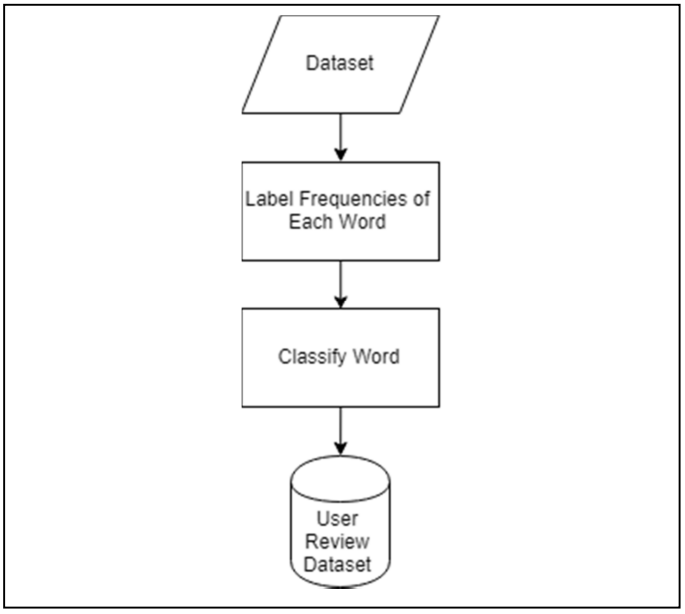


Fig. 3. Stopwords Construction Diagram

Stopwords is a meaningless word [7]. The presence of Stopwords makes the classification has a lower accuracy. By removing Stopwords can improve the final result. In this study, we compare the two Stopwords. That is the General Stopwords and the Context-based Stopwords. General Stopwords are obtained from existing TF-IDF functions. While Context-based Stopwords are obtained by searching for positive and negative ratios within the dataset.

Fig. 3 explains how Stopwords are formed. First, each dataset calculates the labeled frequency for each word. For example, can be seen in Table 1. Column words contain word list, column positives contain positive label frequency, column negative contains negative label frequency, column total contains the frequency of the word appears in the dataset.

TABLE I
DATASET LABEL FREQUENCY

No	Words	Positive	Negative	Total
0	Wow	2	0	2
1	...	20	19	39
2	Loved	2	0	2
3	this	48	48	96
4	place	56	49	105
5	.	387	434	821
6	Crust	0	1	1
7	is	105	69	174
8	not	17	89	106

Second, classify the word is it a stopword or not. One way that can be used is by finding the ratio of positive frequencies to the total. Ratio positive is the result of the positive frequency divided by total frequency as can be seen in Equation (1) below. Then we can conclude the word stopword or not in the Equation (2) below. For example, the word "is". The word "is" cannot be used to distinguish a sentiment. It is Actually more difficult because it has an ambiguous context.

$$Ratio_{positive} = \frac{Positive}{Total} \quad (1)$$

$$Stopwords\ \{IF\ Ratio_{positive} \geq 0.4\ AND\ Ratio_{positive} \leq 0.6\} \quad (2)$$

Lastly, Stopwords are saved into the array according to the initial dataset.

C. Features Construction

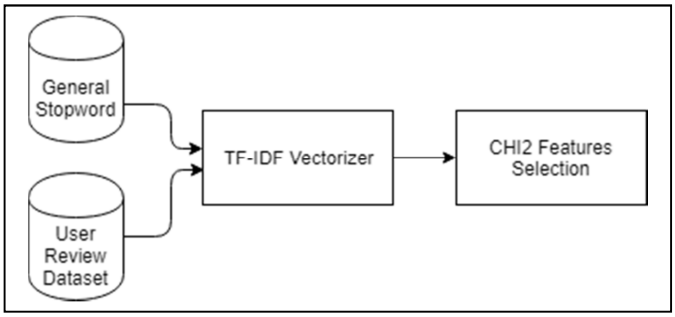


Fig. 4. Features Construction Diagram with General Stopwords

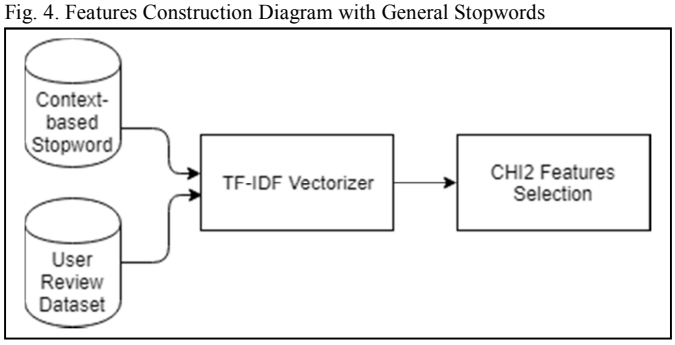


Fig. 5. Features Construction Diagram with Context-based Stopwords

In this section, we focus on generating features from a text and optimizing it. We used TF-IDF to generate features and we use Stopwords and CHI2 to optimize the feature. In this section, we use 2 different methods to compare. There is the use of Stopwords on TF-IDF and Context-based Stopwords.

TF-IDF or called term frequency-inverse document frequency is an efficient approach for determining the weight for each feature [12]. We used the TF-IDF Vectorizer Sklearn method in python to makes features in this study. The result of the features can be huge and may contain noisy and irrelevant features.

In this study, to make the feature filtered from the noise and irrelevant features we use 2 methods. There are Stopwords removal and CHI2 features selection. There are two Stopwords that will be compared. General Stopwords from TF-IDF (Fig. 4) and Context-based Stopwords (Fig. 5). The feature that contains Stopwords will be removed.

Besides using Stopwords Removal. In this study, we also use CHI2 features selection. Which is popular for sentiment analysis features selection. The CHI2 value represents the association between the text feature and its associated sentiment class [4], [13]. We used the CHI2 Features Selection Sklearn method in python to makes features in this study. We also compare the number of features used to determine the optimal number of features.

D. Classification



Fig. 6. Classification Diagram

In this section, we receive input in the form of features that have been fixed from the previous process and produce output in the form of predictable sentiment label from data test. That can be look in Fig.6. Machine learning that we use in doing data train and data test is Gaussian Naive Bayes. We used K-folds Cross-validation in order to obtain valid results.

Cross-validation is a statistical method that can be used to evaluate the performance of a model or algorithm in which data is separated into two subsets data train and data test. Data train is used for data training to build the model and the data test is used for data testing for validating the data. One of the Cross-validation is K-fold. K-fold is usually used because it can reduce computational time while maintaining the accuracy of the estimation. 10-fold Cross-validation is the best model selection, it tends to provide a less biased estimation of the accuracy [10].

Machine learning algorithm that we used for data training and data testing is Gaussian Naïve Bayes. The ratio of data train and data test in this paper is 9:1. Naïve Bayes is a popular machine learning for classifying. Many domains have minim misclassification rate [11].

E. Comparing Results

In this study, we compared the results of each dataset used. In addition, we compared the results from the minimal number features of CHI2 feature selections and compared the use of Stopwords on TF-IDF with Stopwords based on the context generated from the dataset. The results to be compared are measured using accuracy, precision, recall, and f1-score. Precision is the ability of a classification model to identify only the relevant data points [8], see Equation (3). Recall is the ability of a model to find all the relevant cases within a dataset [8], see Equation (4). F1-score is the harmonic mean of Precision and Recall [8], taking both metrics into account in the Equation (5). Accuracy is the quality or state of being correct or precise [8], see Equation (6).

$$Precision = \frac{TP}{(TP+FP)} \quad (3)$$

$$Recall = \frac{TP}{(TP+FN)} \quad (4)$$

$$F1Score = 2\,x\,\frac{(Precision\ x\ Recall)}{(Precision + Recall)} \quad (5)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

IV. DISCUSSION AND RESULT

I build a python program that has an input CSV file that contains the dataset of the review. The dataset contains comment and sentiment for each review. We took user review from UCI machine learning repository (https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences) [9]. The dataset contains 500 positive and 500 negative sentences and sentence score is either 1(for positive) or 1(for negative). The result of the experimental study can be seen in Table 2.

TABLE II. EXPERIMENTAL RESULT

Dataset	Number of Features		Number of features best (CHI2)	Precision			Recall			accuracy			f1-scores		
	general stopwords	context-based Stopwords		general stopwords	context-based Stopwords	Diffrence	general stopwords	context-based Stopwords	Diffrence	general stopwords	context-based Stopwords	Diffrence	general stopwords	context-based Stopwords	Diffrence
Yelp.com	884	1078	5%	0.6	0.66	0.06	0.99	0.93	-0.06	0.68	0.72	0.04	0.72	0.73	0.01
			10%	0.64	0.66	0.02	0.98	0.98	0	0.75	0.77	0.02	0.75	0.77	0.02
			15%	0.67	0.7	0.03	0.99	0.98	-0.01	0.78	0.81	0.03	0.77	0.79	0.02
			20%	0.69	0.72	0.03	0.99	0.98	-0.01	0.8	0.83	0.03	0.79	0.8	0.01
			25%	0.71	0.77	0.06	0.99	0.97	-0.02	0.82	0.87	0.05	0.8	0.83	0.03
			30%	0.73	0.77	0.04	0.99	0.97	-0.02	0.85	0.87	0.02	0.82	0.84	0.02
			35%	0.75	0.79	0.04	0.99	0.97	-0.02	0.86	0.88	0.02	0.83	0.85	0.02
			40%	0.76	0.8	0.04	0.99	0.98	-0.01	0.88	0.89	0.01	0.84	0.86	0.02
			45%	0.77	0.8	0.03	0.99	0.98	-0.01	0.88	0.9	0.02	0.85	0.86	0.01
			50%	0.78	0.8	0.02	0.99	0.99	0	0.89	0.9	0.01	0.86	0.86	0
			55%	0.82	0.8	-0.02	0.99	0.99	0	0.91	0.9	-0.01	0.89	0.86	-0.03
			60%	0.78	0.8	0.02	0.98	0.99	0.01	0.88	0.9	0.02	0.85	0.86	0.01
			65%	0.78	0.8	0.02	0.97	0.98	0.01	0.88	0.9	0.02	0.85	0.86	0.01
			70%	0.77	0.8	0.03	0.94	0.98	0.04	0.86	0.9	0.04	0.84	0.86	0.02
			75%	0.76	0.79	0.03	0.92	0.97	0.05	0.84	0.89	0.05	0.82	0.85	0.03
			80%	0.75	0.79	0.04	0.89	0.95	0.06	0.82	0.88	0.06	0.8	0.84	0.04
			85%	0.72	0.78	0.06	0.84	0.94	0.1	0.79	0.87	0.08	0.76	0.83	0.07
			90%	0.69	0.77	0.08	0.81	0.92	0.11	0.76	0.85	0.09	0.73	0.81	0.08
			95%	0.68	0.76	0.08	0.79	0.88	0.09	0.74	0.82	0.08	0.71	0.79	0.08
			100%	0.66	0.74	0.08	0.77	0.85	0.08	0.71	0.79	0.08	0.69	0.76	0.07
Imdb.com	1186	1298	5%	0.6	0.71	0.11	0.99	0.85	-0.14	0.68	0.74	0.06	0.74	0.74	0
			10%	0.65	0.86	0.21	0.98	0.71	-0.27	0.75	0.8	0.05	0.78	0.77	-0.01
			15%	0.71	0.71	0	0.94	0.97	0.03	0.78	0.8	0.02	0.8	0.81	0.01
			20%	0.71	0.76	0.05	0.97	0.96	-0.01	0.8	0.84	0.04	0.82	0.84	0.02
			25%	0.74	0.96	0.22	0.97	0.79	-0.18	0.82	0.88	0.06	0.84	0.87	0.03
			30%	0.75	0.96	0.21	0.98	0.78	-0.2	0.85	0.87	0.02	0.85	0.86	0.01
			35%	0.77	0.96	0.19	0.98	0.83	-0.15	0.86	0.9	0.04	0.86	0.89	0.03
			40%	0.79	0.97	0.18	0.98	0.81	-0.17	0.88	0.89	0.01	0.87	0.88	0.01
			45%	0.8	0.98	0.18	0.98	0.81	-0.17	0.88	0.89	0.01	0.88	0.88	0
			50%	0.82	0.97	0.15	0.98	0.82	-0.16	0.89	0.9	0.01	0.89	0.89	0
			55%	0.82	0.98	0.16	0.98	0.83	-0.15	0.91	0.9	-0.01	0.89	0.9	0.01
			60%	0.83	0.98	0.15	0.98	0.84	-0.14	0.88	0.91	0.03	0.89	0.9	0.01
			65%	0.82	0.98	0.16	0.97	0.84	-0.13	0.88	0.91	0.03	0.89	0.9	0.01
			70%	0.81	0.97	0.16	0.97	0.82	-0.15	0.86	0.9	0.04	0.88	0.89	0.01
			75%	0.81	0.97	0.16	0.94	0.82	-0.12	0.84	0.9	0.06	0.87	0.89	0.02
			80%	0.79	0.94	0.15	0.91	0.81	-0.1	0.82	0.88	0.06	0.85	0.87	0.02
			85%	0.77	0.91	0.14	0.89	0.8	-0.09	0.79	0.87	0.08	0.83	0.85	0.02
			90%	0.76	0.88	0.12	0.87	0.8	-0.07	0.76	0.85	0.09	0.81	0.83	0.02
			95%	0.75	0.85	0.1	0.85	0.78	-0.07	0.74	0.83	0.09	0.79	0.81	0.02
			100%	0.73	0.82	0.09	0.82	0.77	-0.05	0.71	0.81	0.1	0.77	0.79	0.02
Amazon.com	826	1023	5%	0.6	0.64	0.04	0.99	0.97	-0.02	0.66	0.71	0.05	0.74	0.77	0.03
			10%	0.65	0.71	0.06	0.98	0.96	-0.02	0.73	0.79	0.06	0.78	0.82	0.04
			15%	0.71	0.85	0.14	0.94	0.88	-0.06	0.77	0.86	0.09	0.8	0.86	0.06
			20%	0.71	0.78	0.07	0.97	0.95	-0.02	0.79	0.85	0.06	0.82	0.86	0.04
			25%	0.74	0.8	0.06	0.97	0.96	-0.01	0.81	0.86	0.05	0.84	0.87	0.03
			30%	0.75	0.81	0.06	0.98	0.97	-0.01	0.83	0.87	0.04	0.85	0.88	0.03
			35%	0.77	0.82	0.05	0.98	0.97	-0.01	0.85	0.88	0.03	0.86	0.88	0.02
			40%	0.79	0.83	0.04	0.98	0.97	-0.01	0.86	0.88	0.02	0.87	0.89	0.02
			45%	0.8	0.84	0.04	0.98	0.97	-0.01	0.87	0.9	0.03	0.88	0.9	0.02
			50%	0.82	0.86	0.04	0.98	0.97	-0.01	0.88	0.91	0.03	0.89	0.91	0.02
			55%	0.82	0.87	0.05	0.98	0.97	-0.01	0.88	0.92	0.04	0.89	0.92	0.03
			60%	0.83	0.88	0.05	0.98	0.98	0	0.89	0.92	0.03	0.89	0.92	0.03
			65%	0.82	0.89	0.07	0.97	0.98	0.01	0.88	0.93	0.05	0.89	0.93	0.04
			70%	0.81	0.87	0.06	0.97	0.98	0.01	0.87	0.92	0.05	0.88	0.92	0.04
			75%	0.81	0.86	0.05	0.94	0.97	0.03	0.86	0.91	0.05	0.87	0.91	0.04
			80%	0.79	0.86	0.07	0.91	0.95	0.04	0.84	0.9	0.06	0.85	0.9	0.05
			85%	0.77	0.84	0.07	0.89	0.93	0.04	0.82	0.88	0.06	0.83	0.88	0.05
			90%	0.76	0.82	0.06	0.87	0.9	0.03	0.8	0.85	0.05	0.81	0.86	0.05
			95%	0.75	0.79	0.04	0.85	0.87	0.02	0.78	0.83	0.05	0.79	0.83	0.04
			100%	0.73	0.78	0.05	0.82	0.85	0.03	0.76	0.81	0.05	0.77	0.81	0.04

TABLE III.
A COMPARED TO CONTEXT-BASED STOPWORDS

Dataset	Number of Features	
	General Stopwords	*Context-based Stopwords*
Yelp	884	1078
IMDb	1186	1298
Amazon	826	1023

Table 3 shows about the number of features for each dataset. General Stopwords produces fewer features than those of Context-based Stopwords. General Stopwords remove more features than those of Context-based Stopwords.

TABLE IV
THE COMPARISON BETWEEN THIS STUDY AND ANOTHER STUDY

Dataset	Average Difference			
	Precision	*Recall*	*F1-Score*	*Accuracy*
Yelp	0.040	0.020	0.025	0.038
IMDb	0.145	-0.125	0.013	0.045
Amazon	0.059	0.001	0.036	0.048
Average	**0.081**	**-0.035**	**0.025**	**0.043**

Context-based Stopwords improves precision 8%, F1-score 2.5%, accuracy 4.3%, and decreases recall 3.5%. The results of the accuracy, precision, recall, and f1 scores are pretty good. In this experiment appears some problems like slang word, dataset number, and number in the sentence cannot be the feature. This result can be seen in Table 4.

However, in the case of IMDb dataset decreased recall score up to 12% and increased precision up to 14%. But, increase the f1-score up to 1.3% and accuracy 4.5%. This means, even though the recall reduces, the accuracy still increases. Table 4 shows the greatest difference average is in the precision score up to 8%.

TABLE V
THE COMPARISON BETWEEN THIS STUDY AND ANOTHER STUDY

Dataset		Methods	
		Context-based Stopwords	*GICF with Embeddings on Sentences [14]*
Accuracy	Yelp	90%	92%
	IMDb	91%	88%
	Amazon	93%	88%

Table 5 shows a comparison between Context-based Stopwords method in this study and GICF with embeddings on Sentence in "From Group to Individual Labels using Deep Features" study by K. Dimitrios, D. Misha, D.F. Nando, and S. Padhraic [14]. The comparison using the same dataset. GICF method using Convolutional Neural Network Deep Learning. Context-based Stopwords methods better than GICF in IMDb and Amazon dataset. In Yelp dataset, GICF has a better result than Context-based Stopwords.

There are three things that can improve accuracy in this study:

A. Slang Word

User review usually did not use standard sentences. For example, "I recommend this for EVERYONE who loves film, movies, anything...A Work of Art!" the word in ", anything...A" cannot be transformed to the exact word. So, to optimize the result we must build a slang word method. That can transform the sentence to "I recommend this for everyone who loves film, movies, anything a work of art".

B. Dataset Number

In this study has 1000 sentences for each dataset. The features that produce from this dataset is not more than 1300. The bigger the dataset and the balance of the dataset can make better results. The bigger dataset can represent all the possibility.

C. The Number in The Sentence Cannot be a Feature

Some of the users put a rating in their review. For example, "1/10 - and only because there is no setting for 0/10." It can be a key of the classifying. "0/10" or "1/10" it means a negative sentiment. If the TF-IDF can detect this to transform it into the vector it will help the results.

V. CONCLUSIONS AND FUTURE WORKS

According to the result of the approach from the dataset, we can conclude that the Supervised Method can be optimized by using CHI2 features selection and Stopwords Removal. The use of Context-based Stopwords enriched the number of Stopwords which removed biased features.

There are three parts to optimize the approach in this paper. The first part makes a function to detect and clean the slang word. The second part increases the dataset number that can represent all the possibility. The third part makes a function to detect user rating in the user review.

REFERENCES

[1] Internetworldstats.com. (2018). World Internet Users Statistics and 2018 World Population Stats. [online] Available at: https://www.internetworldstats.com/stats.htm [Accessed 6 May 2018].

[2] R. Rohini S and M. Emmanuel, "Feature Selection for Sentiment Analysis by using SVM," Vol.84(5), December 2013, pp. 21-25.

[3] S. ChandraKala1 and C. Sindhu2, "OPINION MINING AND SENTIMENT CLASSIFICATION: A SURVEY," Vol.3(1), October 2012, pp. 420-427.

[4] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," IEEE Artificial Intelligence, 2017. doi:10.1109/tai.1995.479783

[5] S. Rani and P. Kumar, "A Sentiment Analysis System to Improve Teaching and Learning," in Computer, vol. 50(5), pp. 36-43, 2017. doi:10.1109/MC.2017.133

[6] Y. M. Aye and S. S. Aung, "Sentiment analysis for reviews of restaurants in Myanmar text," IEEE Software Engineering. Artificial Intelligence. Networking and Parallel/Distributed Computing (SNPD). Japan, pp. 321-326, 2017. doi:10.1109/SNPD.2017.8022740

[7] H. Saif, M. Fernandez, Y. He, and H. Alani, "SentiCircles for Contextual and Conceptual Semantic Sentiment Analysis of Twitter, " Lecture Notes

in Computer Science The Semantic Web: Trends and Challenges, 2014, pp. 83-98. doi:10.1007/978-3-319-07443-6_7

[8] M. Jupri and R. Sarno, "Taxpayer compliance classification using C4.5, SVM, KNN, Naive Bayes and MLP," in 2018 International Conference on Information and Communications Technology (ICOIACT), 2018, pp. 297–303. DOI: 10.1109/ICOIACT.2018.8350710

[9] D. Dua and E.T. Karra. (2017). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[10] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," In Proceedings of International Joint Conference on AI (IJCAI), 1995, pp. 1137–1145.

[11] P. Domingos and M. Pazzani, Machine Learning (1997) 29: 103. DOI:10.1023/A:1007413511361

[12] B.Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," International Conference on Data and Software Engineering (ICoDSE), 2015, pp. 170 – 174. DOI: 10.1109/ICODSE.2015.7436992.

[13] L. Yang, B. Jian-Wu, and F. Zhi-Ping. (2017). Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms. Expert Systems with Applications. 80. 10.1016/j.eswa.2017.03.0

[14] K. Dimitrios, D. Misha, D.F. Nando, and S. Padhraic, "From Group to Individual Labels using Deep Features," in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015, pp.597-606.

[15] B.S. Rintyarna, R. Sarno, and C. Fatichah, "Enhancing the performance of sentiment analysis task on product reviews by handling both local and global context," International Journal of Information and Decision Sciences. 2018 Vol. 11.

[16] F.H. Rachman, R. Sarno, and C. Fatichah, "Music Emotion Classification based on Lyrics-Audio using Corpus Based Emotion," International Journal of Electrical and Computer Engineering (IJECE), vol. 8, no. 3, pp. 1720-1730, Juni 2018. DOI: 10.11591/ijece.v8i3.pp1720-1730

[17] E.W. Pamungkas, R. Sarno, and A. Munif, "B-BabelNet: Business-Specific Lexical Database for Improving Semantic Analysis of Business Process Models," Telecommunication, Computing, Electronics and Control (TELKOMNIKA), Vol 15, No 1, 2017. DOI: http://dx.doi.org/10.12928/telkomnika.v15i1.3176

[18] F.H. Rachman, R. Sarno, and C. Fatichah. "CBE : Corpus-Based off Emotion for Emotion Detection in Text Document," The 3rd International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE), Pp. 331 – 335, 2016. DOI: 10.1109/ICITACEE.2016.7892466

[19] B. Rintyarna and R. Sarno, "Adapted Weighted Graph for Word Sense Disambiguation," The 4th International Conference on Information and Communication Technology (ICoICT), 2016. DOI: 10.1109/ICoICT.2016.7571884

Improving The Accuracy of Naïve Bayes Algorithm for Hoax Classification Using Particle Swarm Optimization

[1]Akhmad Pandhu Wijaya, [2]Heru Agus Santoso
Computer science
Dian Nuswantoro University
Email : wijayapandhu39@gmail.com, heru.agus.santoso@dsn.dinus.ac.id.

Abstract— **Hoax news circulation is very widespread which occurs in the media information, both print and online media. For some people hoax news can only appear on the online media. But printed medias also often include hoax news in their published news. In the present era, it is very important providing information with relevant and additional facts otherwise it is categorized as hoax. Therefore, hoax classification approach is needed. This paper focuses on improving the accuracy of hoax classification in textual documents contents. Naive Bayes algorithm is used to train dataset with the use of PSO in the algorithm. Experiment is conducted with the trained model over 600 documents. It shows that feature selection with PSO affects the classification results performed using Naïve Bayes. Accuracy increased from 91.17% without using feature selection, to 92.33% when feature selection is carried out using PSO.**

Keywords— text classification, hoax, news, PSO, Naïve Bayes.

I. INTRODUCTION

Rising news circulation of hoax in various media, both online media and printed news media become a serious problem for social life. Some peoples think that hoax news may only appear on online media, but in fact printed media often share hoax news on their news column. However, hoax news with hate speech content mostly appear in social media. The present era is a reflection of the concerns in the history of human life. With technological advances and the ease of access to information, it makes people very readily accept information and trust eventhough the information may be categorized as hoax with hate speech contents. There are many cases happen in our society related to hate speech, as well as events with facts that do not occur. But as if happening and packed with patterns as best as possible, then it aimed at fostering the spirit so that people believe it because of sedition.

If it is viewed in terms of science and technology, today's era is very important to make improvements of hoax classification problem, especially in the classification of text. With the classification of all text contents, text documents can be grouped based on their parameters. We want to do so to facilitate in obtaining a better structured information for reader. However, there are some technical problems in the classification of them, e.g., on the feature selection section. It can be said that all of the features are not appropriate and can be used as a basis to build a classification model. In fact, not all features can be used as a parameter for hoax classification. To select features that match our expectations and to reduce features that do not fit in the feature selection process, it becomes so important. That is why, this paper focuses on the feature selection process. Beside the society must begin to be educated to realize the danger of news hate speech. With the availability of data and appropriate algorithm, this study hopes able to classify news into decent and organized information in a better manner.

This study aims to develop classification pattern of news hate speech, in the hope that it will improve performance of the existing research. Our proposed approach can make optimal synergy in handling the problems caused by the news as it happens nowadays. While in the field of education that is triggering more complex thoughts in the field of assessment of news hoax, so as to make the intellectuals more concerned with the problems that exist in the community. Especially in the development of news hate speech classification and able to realize human civilization for a better society. Improving the ability of hoax classification is carried out by conducting research in the field of text mining empowered with natural language processing based on the urgent need for improving its performance result.

II. RESEARCH METHODOLOGY

Study that attempts to search or extract data pattern from a collectoion of data from various types resources is called mining. While text mining is extracting data pattern in the form of text that aims to define a set of knowledge. Figure 1 depicts text mining stages which mainly consist of four processes.

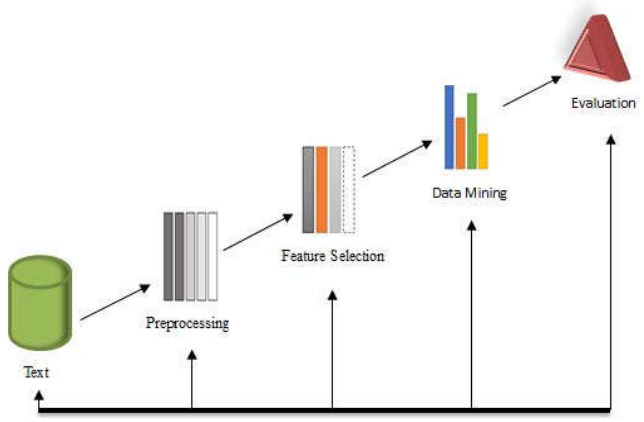


Figure 1 Text Mining Stage

These processes have their respective schemes interconnected with each other because in the results of each process performed will be the main ingredients for the next process. In the text shows that the data that has been collected into a collection of data with the type of text and unknown structures that exist in it. Preprocessing stage will clean up the original document contents which consists of various kinds of characters until the text is ready to be processed. After that, the data becomes more structured as it is necessary in the efforts to clarify the existing documents. On the stage of feature selection, existing features need to be selected in the process of classification into desired class.

2.1 The meaning of text

Indonesian Language has a standard reference in writing that becomes the signs for governance writing. So that it becomes a good communication method. Sentences written on the news as well as information consist of a small section that compiles the word. Basically the word is a unit that is able to stand alone, with one word we can understand what is meant but it is very weak and vulnerable to the question of what the meaning of the word expressed. To express a more complex thought requires the union of words summarized in the sentence, so as to bring up the meaning in accordance with what is meant in the information. To find a sentence and process it with a classification model, clustering and other algorithms, s set of appropriate words is needed that basically has greater impact in the meaning of the sentence or information.

2.2 Hate Speech

The study in this article is centered on hoax news with hate speech content that has a distinctly different pattern of news writing in general. As news hoax often have irregular patterns with inappropriate language, provocative or overstatement and so forth, some of the research that suggested hoax news taken from the media communication either email, news, or chat [9]. The hoax news itself in the dictionary has meanings of jokes, false stories, delinquents, jokes, lies, cheats and so on, and hoaxes have synonyms (practical joke, joke, jest, prank, trick). While in Kamus Besar Bahasa Indonesia (KBBI), hoax is defined as "false news", then hoax can be translated with simple understanding.

2.3 Text Preprocessing

In the text preprocessing stage, cleaning unimportant characters is performed in the pre-processing stage, This stage processes the data until it becomes ready and meaningful. In general, documents have diverse characteristics and have various deimensions, the characteristics of which are meant documents with nois or disturbing parts, and the layout of the text on the documents are unstructured. The structure of the document becomes very important to introduce the pattern so as to characterize a particular class. At this stage there are several steps that must be carried out such as case folding, tokenizing, filtering, and stemming.

a. Case Folding

The consistency of writing documents using capital letters is not always met. Then the capital letter used must be equated with eliminating all the writing with a capital letter by case folding process as depicted in Figure 4. As in the word "TESIS", "Thesis", or "TeSis, when they are input with similar writing model, they will still be read as " thesis ".

b. Tokenizing

The tokenizing process is where text data in the form of paragraphs, sentences, or words are processed with the aim of dividing into specific parts. It can be said they will be converted in to a set of unit of words. Tokezing terminology is a process of replacing the value of the original value into a value called "token" [10]. If the dataset contains a sentence "classification text on hoax news document" it will generate a set of words consits of "classification", "text", "on", "document", "news", "hoax". That way the data that was composed into a paragraph or

sentence will be the units of words that will be the element of document processing. Benefits of tokenizing is simplify the process of classification, with separation of the word elements into the input material becomes more concise. The example of tokenizing process is depicted in Figure 5

c. Filtering

In this section is often also called a stopword list or stop word removal. And this process is a standard operation which is performed on the data in the form of text that aims to take the words result of tokenizing. It then become a representation of documents that often called as a set of features. Technically, the processing using the stop list model allows the user to create a word dictionary as the basis of disposing of unused words on the dataset. Figure 6 and Figure 7 descibe the process of filtering.

ada	apabila	baik	benarkah	berkenaan
adalah	apakah	bakal	benarlah	berlainan
adanya	apalagi	bakalan	berada	berlalu
adapun	apatah	balik	berakhir	berlangsung
agak	artinya	banyak	berakhirlah	berlebihan
agaknya	asal	bapak	berakhirnya	bermacam
agar	asalkan	baru	berapa	bermacam-macam
akan	atas	bawah	berapakah	bermaksud
akankah	atau	beberapa	berapalah	bermula
akhir	ataukah	begini	berapapun	bersama
akhiri	ataupun	beginian	berarti	bersama-sama
akhirnya	awal	beginikah	berawal	bersiap
aku	awalnya	beginilah	berbagai	bersiap-siap
akulah	bagai	begitu	berdatangan	bertanya
amat	bagaikan	begitukah	beri	bertanya-tanya
amatlah	bagaimana	begitulah	berikan	berturut
anda	bagaimanakah	begitupun	berikut	berturut-turut
andalah	bagaimanapun	bekerja	berikutnya	bertutur
antar	bagi	belakang	berjumlah	berujar
antara	bagian	belakangan	berkali-kali	berupa
antaranya	bahkan	belum	berkata	besar
apa	bahwa	belumlah	berkehendak	betul
apaan	bahwasanya	benar	berkeinginan	betulkah

Figure 2 a set of features

d. Weighting

Large-scale documents need to be processed with the appropriate scheme to generate each weight of the object in the form of words, phrases or units in a document that has been processed using indexing. With the process of indexing, it makes the data to have a priority scale based on the numbers of obtained words or features in accordance with their weight that has been determined.

1. Term Frequency (TF)
   The document has a term that will be the determinant of classification. This term is a vector format that can be understood by the computer and the vector frequency is very influential [11]. The frequency of occurrence of the term is declared with Term Frequency (TF). With frequency, set of terms exist in documents can be analyzed with the parameters of the number of terms that appear. So that the value of the document adjustment depends on the larger weights. Term Frequency has several categories that are used as a reference to formulate such a model:
2. Binary TF
   Focused only on the existence of a term or a word on the document, whether the term is contained in the document or not. If the term is on the document, it will be assessed with (1), otherwise it will be given a value (0).
3. Raw TF
   The number of occurrences of a term in a particular document becomes an existing value in TF-IDF. The value will appear if for example there is one term appearing 10 times in the document then the value obtained is 10.
4. Log TF
   Existing documents and slightly contain the desired terms need to be selected so that the frequency of the terms in document decreases.
5. Normalized TF
   It is the ratio of the number of terms / frequency of a term to the frequency of all terms in the document.

The proposed approach is formulated to calculate the value of TF and IDF on every word. The weight calculation on each token in the document is spelled out in the formula :

$$W_{dt} = tf_{dt} * IDF_t \quad (1)$$

5. Inverse Document Frequency (IDF)
   It is the calculations performed to reduce the weight of a term scattered across most documents and make the search process inhibited. The parameter that appears is the less number of documents containing a certain terms will has greater its idf value.

$$IDF_j = \log\left(\frac{D}{df_t}\right) \quad (2)$$

Variables in the formula (1) and formula (2) are explained as follows:

W_{dt} : The word weights d in document t

t_{dt} : The number of occurrences of the word d in document t

D : Number of documents

df_t : Number of documents t containing the word d.

2.4 Feature Selection

Categorization with features in large numbers of collection of documents has its own difficulties in text classification. High feature dimension is very influential in the process of classification because there may exist duplication and redundancy of the feature. The PSO method is an approach that runs on the basis of related interaction patterns and takes into account the quantity factor in the form of population as the main basis of research with large data. This model was initiated by Kennedy and Eberhart in 1995 [2]. The basis of the PSO concept is to represent the problem as a particle present in a collection (swarm), vector is represented as $x_i = (x_{i1}, x_{i2}, \dots, x_{iD})$, where D is a measure that reflects the search space and then the space is entered by the particle that will find the best solution from the search on that variable. The particles in question can be represented as :

$$v_i = (v_{i1}, v_{i2}, \dots, v_{iD}).$$

3. Classification

The process for defining models with different concepts or data classes to provide data estimates refers to the available classes. This process is done after the preprocessing step. The working concept of Naïve Bayes Classifier is with probability on the dataset. Then, machine classifies with this model to determines the probability of the document against a particular class, illustrated that a document D associated with the given class C is $P(C|D)$ [3][4].

$$P(C|D) = \frac{P(D|C)\,P(C)}{P(D)} \quad (3)$$

Where:

P = Probability
C = Class
D = Document

The probability of a document against a given class depends on several variables that become benchmarks, i.e., class probability to document, class probability, and document probability.

4 Evaluation

It is necessary to test and evaluate the proposed model, to know the success rate. The evaluation metrics used in this study are: accuracy, recall, and precision. The value of each metrics are obtained form the so called confusion matrix which is produced during classification process. From the proposed evaluation model, then there will be parameters to know the success, but it is not possible that the results do not match the expectations because the results obtained depends on several factors in it. While the calculation of precission uses the number parameters.

III. DISCUSSION

Feature selection approach design consists of training and testing set. The start-up phase or the training of data is done with preprocessing process which produces the data in the form of features and frequency. Based on the weight generated by preprocessing in the form of weighted values for later selection of features using the PSO, then the resulting feature will be ready to do further process of classification.

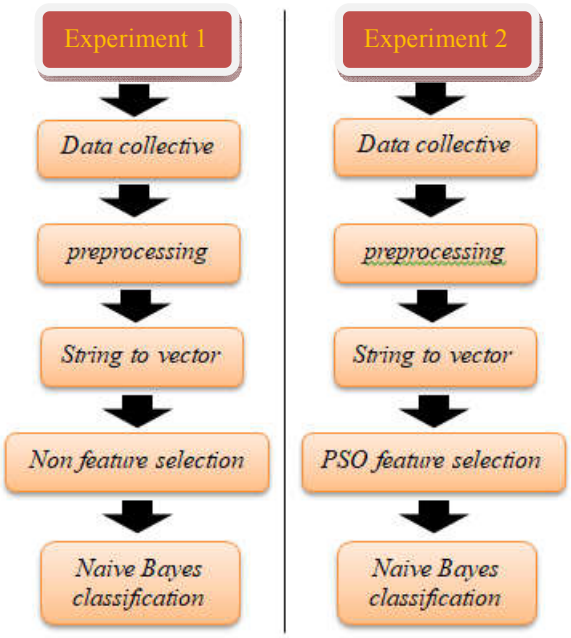


Figure 3 Research Steps

3.1 Data Collection

All of hate speech hoaxes and real data are collected from trusted sources. For the hoax category, the dataset collection is taken from a portal which is specifically acknowledged to present hoax information i.e., https://www.turnbackhoax.id. The data is supplemented with supporting data in the form of text from social media containing hate speech, which is taken from research from Uiversitas Indonesia [13]. The study used data from twitter containing hate speech, as shown below.


NARASI
"Satu lagi leting kita putra terbaik bangsa..alumni SEBA POLRI SINGARAJA (SEBAS) 1986/1987 PLETON I jadi calon PANGLIMA TNI.BRAVO SEBA V SINGARAJA.baru setahun tugas di res sumba ntt. lulus AKABRI AU.".

Figure 4 example of hoax dataset

JAKARTA - Setelah stagnan di Februari, pemerintah mengumumkan perubahan harga bahan bakar minyak (BBM)
Sedangkan harga minyak tanah masih Rp 2.500 per liter, begitu juga dengan solar yang tetap dijual Rp 6
Kepala Pusat Komunikasi Publik Saleh Abdurrahman mengatakan, kenaikan dikarenakan harga minyak di pasa
Jadi, harus ada penyesuaian untuk harga premium. "Demi menjaga kestabilan sosial ekonomi pengelolaan hi
Terhitung sejak 1 Maret 2015 pukul 00.00 WIB, untuk premium RON 88 di wilayah luar Jawa, Madura, Bali
Harga sebelumnya Rp 6.600 per liter naik menjadi Rp 6.800 per liter.

Untuk daerah Jamali, Direktur Pemasaran Pertamina Ahmad Bambang menyebut, premium yang sebelumnya Rp 6
Namun, seperti biasa, harga di Pulau Bali biasanya lebih mahal karena pajak yang diterapkan lebih tingg
"Kami memahami kondisi masyarakat. Harusnya naik Rp 400 (jadi Rp 7.100, red)," katanya.
Saleh menyebut, harga MoPS Premium mengalami kenaikan menjadi USD 55-70 per barel.
"Kenaikan MoPS sepanjang Februari sebenarnya cukup signifikan. Tapi, Pemerintah tidak menaikkan harga
Untuk solar, harga MoPS sepanjang Februari antara USD 62-74 per barel. Ketidakstabilan harga, lanjut Si
juga terkait pertentangan pelaku pasar minyak dalam menyikapi konflik di Libia. (dim/dio)

Figure 5 example of real dataset

3.2 Preprocessing Dataset

Data processing begins with the process of breaking the data structure into small particles / sentences from the original data with large populations. Then, they are transformed into vectors with different value on each feature, because the frequency that appears different in number. At the beginning of this experiment, an event that appears as a difference in the number of features appears, because this feature is uneven.

Table 1 Example of hoax hate speech

Authentic hoax document
'*Valid* Sangat luar biasa zholim….! Masih ingat musibah Crane di Masjidil Haram..? Kerajaan Saudi bertanggung jawab dan akan menyantuni para korban, tnyata sudah cair sejak lama tapi belum di salurkan kepada korban oleh KEMLU , dengan las an masih verifikasi…! Brp lama verifikasinya pak ..? Sampai kapan itu diverifikasi? Sementara Kafir Harbi , Syiah dan kelompok munafik terus menerus membuat fitnah kepada Arab Saudi. Bantu Share…!
After processed
valid sangat luar biasa zholim masih ingat musibah crane di masjidil haram kerajaan saudi bertanggung jawab dan akan menyantuni para korban tnyata sudah cair sejak lama tapi belum di salurkan kepada korban oleh kemlu dengan alasan masih verifikasi brp lama verifikasinya pak sampai kapan itu diverifikasi sementara kafir harbi syiah dan kelompok munafik terus menerus membuat fitnah kepada arab saudi bantu share

Above table shows that a change of data structure during preprocessing, to get the structure as above. The beginning process are tokenizing and case folding. After the processes are done, then the data must be changed in the form of a matrix to be further processed as presented in Table 2.

Table 2 Matrix of feature weights

Fit/doc	Doc1	Doc2	Doc3	Doc4	Doc5	Doc6	Doc7	Doc8	Doc9	Doc10	Doc11 ?
abua	0	0	0	6,64	0	0	0	0	0	0	
aburizal	1,73	0	3,47	0	3,47	0	0	0	0	0	
aceh	0	0	0	0	0	0	0	9,96	0	0	
adil	4,64	0	0	0	0	2,32	0	0	0	0	
administrasi	0	0	0	3,32	0	0	0	0	0	0	
adu	0	0	0	0	0	0	0	0	0	3,32	
adzab	0	0	0	0	0	0	0	0	0	3,32	
aneh	0	0	0	0	0	3,32	0	0	0	0	
apalagi	0	0	0	0	0	0	6,64	0	0	0	
firaun	0	0	0	0	0	0	0	0	3,32	0	
Kelas	real	real	real	real	real	hoax	hoax	hoax	hoax	hoax	???

3.3 Results Analysis

a. Beginning with the initialization of the dataset and calculating the entire dataset, the number of hoax and real datasets is explained below.

Number of datasets : 600
Number of hoax data : 300
The number of real data : 300

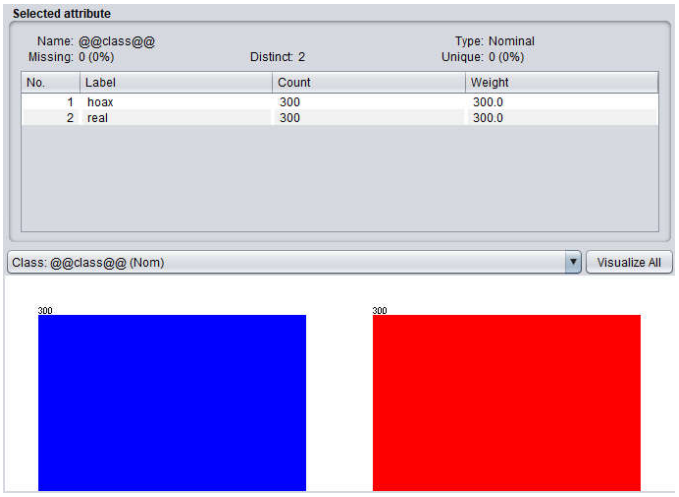


Figure 6 Kind of Class

b. Calculation of weights on the training dataset is collected by documents, which is an equal distribution of each class. Features are selected as a fraction of the existing set of datasets.

c. Naive Bayes experiment with or without PSO are done. Based on two types of experiments above, which are Naive Bayes algorithm without and with feature selection using PSO, the result is promising. The experiment is conducted over 600 documents. It shows that feature selection with PSO improve the classification results performed using Naïve Bayes. The performance increased from 91.17% to 92.33% of accuracy when feature selection is carried out using PSO as depicted in Figure 16.

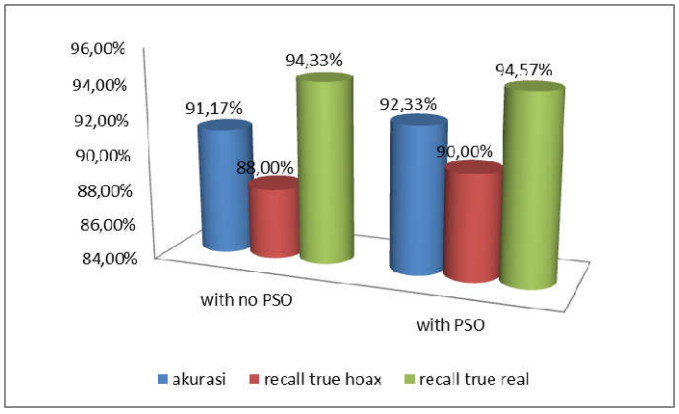


Figure 7 Improved accuracy

IV. RESULT

A. Conclusion

From the discussion as presented in the previous chapter, the authors can draw the following conclusions. Research that has been done and tested, it can be concluded that text feature selection can be done by using Particle Swarm Optimization algorithm and continued by Naïve Bayes classification process. The result indicates that the performance of hoax classification improved by using feature selection.

B. Suggestions

The suggestions that may be given in this study for further development in order to improve the quality and functionality of this document classification method are as follows: Improved text processing and identification and developed the preprocessing stage of the dataset by selecting more text features that are deemed necessary in the dataset to improve the news classification process.

THANK-YOU NOTE

We would like to thank the Government of Indonesia for the funding in the publication of this preliminary study, and also the referees for useful and constructive suggestions.

REFERENCES

[1] J. Vashishtha, "Particle Swarm Optimization based Feature Selection," vol. 146, no. 6, pp. 11–17, 2016.

[2] B. Xue, M. Zhang, and W. N. Browne, "Particle swarm optimization for feature selection in classification: A multi-objective approach," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1656–1671, 2013.

[3] H. Shimodaira, "Text Classification using Naive Bayes," no. 4, 2015.

[4] S. L. Ting, W. H. Ip, and A. H. C. Tsang, "Is Naïve bayes a good classifier for document classification?," *Int. J. Softw. Eng. its Appl.*, vol. 5, no. 3, pp. 37–46, 2011.

[5] M. Vuković, K. Pripužić, and H. Belani, "An intelligent automatic hoax detection system," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5711 LNAI, no. PART 1, pp. 318–325, 2009.

[6] J. C. Hernandez, C. J. Hernandez, J. M. Sierra, and A. Ribagorda, "A first step towards automatic hoax detection," *Proceedings. 36th Annu. 2002 Int. Carnahan Conf. Secur. Technol.*, pp. 102–114, 2002.

[7] J. Meng, H. Lin, and Y. Yu, "A two-stage feature selection method for text categorization," *Comput. Math. with Appl.*, vol. 62, no. 7, pp. 2793–2800, 2011.

[8] Y. Li, A. Algarni, M. Albathan, Y. Shen, and M. A. Bijaksana, "Relevance feature discovery for text mining," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 6, pp. 1656–1669, 2015.

[9] E. Rasywir and A. Purwarianti, "Eksperimen pada Sistem Klasifikasi Berita Hoax Berbahasa Indonesia Berbasis Pembelajaran Mesin," *J. Cybermatika*, vol. 3, no. 2, pp. 1–8, 2015.

[10] R. Tokens, "Tokenization Product Security Guidelines – Irreversible and Reversible Tokens," no. April, pp. 1–84, 2015.

[11] M. Liu and J. Yang, "An improvement of TFIDF weighting in text categorization," *Int. Conf. Comput. Technol. Sci.*, vol. 47, no. Iccts, pp. 44–47, 2012.

[12] I. Ahmad, "Feature Selection Using Particle Swarm Optimization," vol. 2015, no. 4, pp. 231–238, 2015.

[13] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, "Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study," *9th Int. Conf. Adv. Comput. Sci. Inf. Syst. (ICACSIS 2017)*, no. October, 2017.

A Comparison of Data Mining Techniques for Suicide Attempt Characteristics Mapping and Prediction

Kunawut Boonkwang
Human-Hardware Interface and Communications (H^2I-Comm) Laboratory,
Department of Computer Science
Faculty of Science, Khon Kaen University
Mueang, Khon Kaen, Thailand
e-mail:kunawut@kkumail.com

Sumonta Kasemvilas
Human-Hardware Interface and Communications (H^2I-Comm) Laboratory,
Department of Computer Science
Faculty of Science, Khon Kaen University
Mueang, Khon Kaen, Thailand
e-mail:sumkas@kku.ac.th

Sakarin Kaewhao
M.D. Psychiatrist Vice director of Khon Kaen Rajanagarindra Psychiatric hospital
Mueang, Khon Kaen, Thailand
e-mail: jvkk_sakarin@hotmail.com

Orapin Youdkang
RN, MSN (Psychiatric and Mental Health Nursing),
Khon Kaen Rajanagarindra Psychiatric hospital
Mueang, Khon Kaen, Thailand
e-mail: orapin63308@gmail.com

Abstract—According to the report by Khon Kaen Rajanagarindra Psychiatric hospital, there are many factors such as poverty, loss, disappointment, alcohol addiction, addiction, chronic physical illness, depression, economic problems and major life changes that cause suicide attempts. However, there are some unsuccessful suicides. These individuals might try to repeatedly commit suicide attempts. In this paper, the aims of this research are to compare ID3, C4.5 and naïve Bayes and predict the characteristics of individuals who have suicidal ideation to repeatedly commit suicide attempts by using data mining techniques. After data preprocessing, three data mining techniques were selected to compare the efficiency of the single of classification models and the ensemble models. The researchers also adopted synthetic minority over-sampling technique (SMOTE) to balance the suicidal ideation of reattempt and no attempt classes. The researchers used 57 attributes from self-harm surveillance report (RP.506S) of Khon Kaen Rajanagarindra Psychiatric hospital. The experimental results of the single method showed that C4.5 is the best model with the highest percentage of correctly classified instances rate of 90.69%, while the ensemble model, voting method showed the best results of 91.26. We found the number of characteristics from C4.5 is 474 patterns. There are people, who have suicidal ideation over 500 in each case is 42 patterns. In 42 patterns, there are 22 patterns that more than 90% accuracy.

Keywords—Repeated suicide attempts; Classification; Ensemble; SMOTE; Data Mining

I. INTRODUCTION

Suicide is the act of killing yourself intentionally. Suicidal behavior is classified according to the effect of the action. For example, when the action causes death, it is called "completed or successful suicide". When the person tries to commit suicide but still survives, it is called "suicide attempt".When the person has the idea to commit suicide, it is called “suicidal ideation”. [1] There are many factors such as poverty, loss, disappointment, alcohol addiction, addiction, chronic physical illness, depression, economic problems and major life changes that cause suicide attempts. [2] Suicide is a problem that must be solved and be prevented because the effects of suicide are the loss of human resources and the economy of the country. Therefore, this issue is a vital problem to study, understand in order to prevent repeated suicide attempts.

Suicide statistics in Thailand by Khon Kaen Rajanagarindra Psychiatric hospital showed a tendency to increase. In 2015, there is the suicide rate of 6.47 per 100,000 population which increases from 2014 that has suicide rate of 6.07 per 100,000 population. The average suicide is 340 people per month or one person to completed suicide every 2 hours. [3] In this study, the data from self-harm surveillance report (RP.506S) of Khon Kaen Rajanagarindra Psychiatric hospital were used to predict the characteristics of suicide attempters. RP.506S was designed by the National Suicide Prevention Center, Khon Kaen Rajanagarindra Psychiatric hospital, Department of Mental Health, Ministry of Public Health in Thai-

land. RP.506S consisted of four parts including patient information, diagnostic information, influencing factor on suicide attempt, and services.

Data mining is a process that deals with a lot of data to find patterns and relationships that are hidden in that data set. Data mining includes data cleaning, data integration, data selection, data transformation and knowledge representation. There are a lot of data mining techniques such as clustering, association rules, neural networks and classification. Data classification is one of the important techniques of knowledge retrieval on a large database to find unsuspected relationships [4]. There are several techniques namely, decision tree, Naïve Bayes and neural network but the decision tree is one of the more powerful techniques that interprets result properly for prediction. The effectiveness of the decision tree classification has been consistently developed, particularly in terms of accuracy and speed of data classification. This research used decision trees, which yield better results than other methods. These results can be improved by adopting SMOTE [5] to balance the data, and then combine the prediction result from several classifiers using an ensemble voting technique [6].

The aims of this study are to compare the efficiency of the single model and the ensemble models of classification techniques which is suitable for data set of suicide attempts (RP.506S) and to find the characteristics of individuals who has suicidal ideation to repeatedly commit suicide attempt.

The study is organized as follows: Section II, the researchers give a brief theory background and related works. Section III, the researchers give explanation of the research methodology and data used in this work. Section IV, the researchers present the experiments and results. The conclusion and future work are in the last section.

II. THEORETICAL BACKGROUNDS AND RELATED WORKS

A. Decision Tree

The decision tree is one of classification models which uses a tree structure to classify data. Each nodes represent attribute of data set. Each branch is represents the condition from data set, and the leaf node represents the class labeled that the decision tree can create properly. Generally, the rule usually consists of "if, the condition, then, and result". For instance, “if sex=male and alcohol=yes then suicidal ideation = yes”. A typical decision tree is shown in Figure 1.

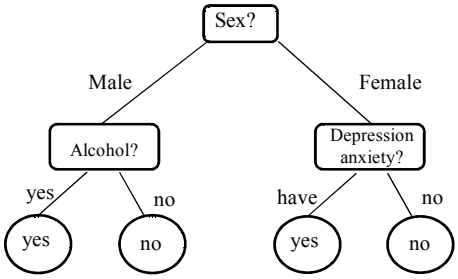


Fig. 1. Example of Decision tree

In figure 1, it represents the pattern of suicide attempt where each branch is divided into two nodes. It predicts characteristics of the person who is repeated suicide attempts in leaf nodes. Inner nodes are shown by rectangle, and leaf nodes are shown by circle. [4]

Decision trees were developed by researchers in order to learn about machine learning, it is called ID3 (Iterative Dichotomiser), in the late 1970s and early 1980s [7]. Later, Ross Quinlan developed the C4.5, which expanded from ID3 for better learning and currently become a new learning standard. [8]

B. SMOTE

The techniques such as decision tree or naïve Bayes are less effective when the information used in the training is not balanced. The imbalance data is the inequality of information in each class which the data in one class is greater than the number of other classes. When imbalanced data is used to classify data, result in the class is more likely to be bias towards other large data groups. [5] SMOTE is an oversampling technique that randomly generates data from a specified number of subclasses. It generates synthesized data from sample data by measuring the distance from the sample point to the nearby data point. The synthesized data is located within the distance from the sample data point to the neighboring data point where the new point is generated.

C. Naïve Bayes

Another classification technique is Bayesian classifiers known as the naïve Bayes, which is a statistical classifier method. Naïve Bayes is a technique used by Bayes' Theorem. Naïve Bayes calculate the probability that is used to predict outcomes when analyzing new cases [4]. Let A is an attribute and C is a class. P (C|A) is the posterior probability that the data with the attribute A has the class C. Similarly, P (A|C) is the probability that the training data contains class C and has the attribute A. P(C) is the prior probability value of class C. Bayes’ Theorem is shown as the following equation:

$$P(C|A)\frac{P(A|C)P(C)}{P(A)} \qquad (1)$$

D. Ensemble Model

Ensemble model are usually used to increase the accuracy of the model. An ensemble is a composite model for classification, derived from a combination of classifiers. In general, the component classifiers tend to be less accurate than ensembles. Bagging, boosting, random forests and voting are examples of popular ensemble model. In Figure 2 shows an ensemble combines a series of k learned models (or base classifiers),$M_1, M_2, \ldots, M_k$, with the purpose of creating an improved composite classification model, M*. [4]

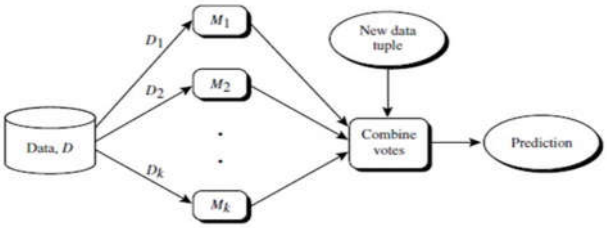


Fig. 2. Ensemble methods

E. Related works

The classification method in data mining has various methods. Decision tree illustrates the separate results of decisions in the form of trees. It is well-known method to solve many problems such as business, education or medical. For solving problems in medical field, Banu and Gomathy [9] used C4.5 (or J48) to predict and help decide the heart disease. Bahrani and Agrawal [10] studied on intestinal carcinoma prognosis of patients to analyze the risk of disease with C4.5 and improved result with SMOTE technique to balance data. They used an ensemble voting with three methods. Likewise, Zhou and Jiang [11] studied on breast cancer prognosis and Wang et al. [6] studied of osteoporosis risk prediction. Both of them used hybrid data mining ensemble using decision tree (C4.5) with artificial neural networks (ANN).

In addition, Xu et al. [12] used ensemble method to analyze the risk of cardiovascular disease. The techniques included naïve bayes, decision tree (J48), neural network, bagging algorithm and support vector machine (SVM). In each technique, the result had different accuracy between 78.98 and 91.48%. However, they used the hybrid method to ensemble four major models namely decision tree, ANN, SVM and bayes network, which made their work more accurate up to 94%. Li et al. [13], Jothi et al. [14], and Rojas et al. [15] used ensemble mining to manage health information with decision tree , swarm intelligence, k-nearest neighbors (KNN), logistic regression, bayesian and SVM. D'monte and Panchal [16] used ensemble to predict the stress level of mental patients with J48, logistic model trees (LMT), randomTree, REPTree, J48 graft, BFTree and Naïve Bayes tree (NBTree). Gosain and Kumar [17] used decision tree together with association rules to analyze health care in HIV infected patients to predict important events in caregivers. It is seen that the decision tree and the ensemble technique are used widespread.

III. Research Methodology

This research presented a method for identifying suicide attempters who are at risk of repeated suicide with a decision tree (C4.5) and to demonstrate the ability of the classification algorithm. First step, data was collected from suicide and self-harm surveillance report (RP.506S) of Khon Kaen Rajanagarindra Psychiatric hospital. Second, data preprocessing was to prepare the information. Data cleaning was used to remove error data and inconsistent data. Next, SMOTE was used to balance the data due to two classes data were imbalance. Then, the researchers compared the performance of the C4.5 model with the ID3 because C4.5 is available for both continuous and discrete data. Next, C4.5 was compared with Naïve Bayes which is a statistical classifier. The ensemble methods were used such as bagging, random forest and voting, which are prevalent ensemble methods. Then these methods were compared their performance. Finally, the experimental results were summarized. All steps were shown in figure 3.

This research used decision trees, which yielded better results than other methods. [9,10] These results can be improved by adopting SMOTE [5] to balance the data, and then combine the prediction result from several classifiers using an ensemble voting technique. [11, 12, 13]

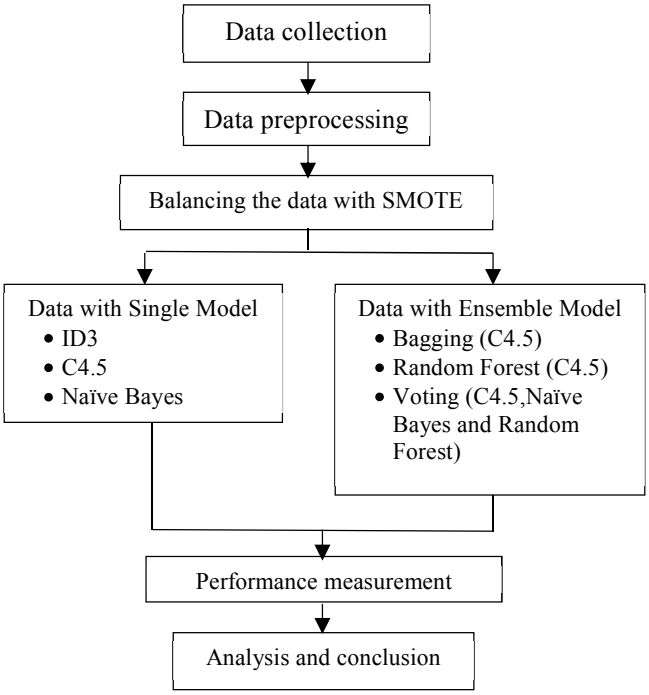


Fig.3. Research Process

A. Data Collection

The data use in this research was collected from the suicide and self-harm surveillance report (RP.506S) of Khon Kaen Rajanagarindra Psychiatric hospital. It was used to collect information of suicide attempters. The data was collected during 2008 to 2016 consisting of 194,489 records and number of items before feature selection is 174 attributes. These attributes are classified into four parts including patient information, diagnostic information, influencing factor on suicide attempt, and services which is shown in Table 1.

Table 1. Details of attribute from RP.506S

No.	Group	Number of attribute	Detail of attribute
1	Patient information	44	Hospital info(8), Informant(8), Name(2), HN(1), Nationality(2), Current address(8), Birthplace(8), Gender(1), Age(1), ID number(1), Marital Status(2), Occupation(2)
2	Diagnosis information	51	Date of service(3), Refer in(2), Type of service(2), Depression anxiety(1), Diagnostic info(24), Self-harm info(7), Method of self-harm(12)
3	Influencing factor on suicide attempt	51	Risk factors for self-harm(27), Health issues info(18), Personal behavior(6)
4	services	28	Service info(14), Admit(1), Death(4), Suicidal ideation(1), Refer out(1), Collector info(7)
Total		174	

B. Data preprocessing

Data preprocessing has the following steps. First, the feature was select by experts' marking who participated in the study. The feature is the key attribute and related data, which is the factor that affected suicide attempt. The number of attribute after feature selection is 57 attributes, which is shown in Table 2. The class can be yes or no of Suicidal ideation. The last data in Table 2 is the preferred answer class.

Table 2. Details of attribute after experts' marking

Patient information		
No.	Attribute	Description
1	Gender	1=Male 2=Female
2	Age	Numbers of age
3	Marital status	1=Single 2=Married 3=Widowed 4=Divorced 5=Others
4	Occupation	1=Agriculture 2=Labors 3=Maid 4=Employees 5=Trading / personal affairs 6=Government / State Enterprises 7=Student 8=Priest 9=Unemployed 10=Others
Diagnosis information		
No.	Attribute	Description
1	Depression anxiety	0=No 1=Yes
Method of self-harm		
No.	Attribute	Description
1	Overdose	0=No 1=Yes
2	Insecticide	0=No 1=Yes
3	Herbicides	0=No 1=Yes
4	Other chemicals	0=No 1=Yes
5	Use of sharp/solid	0=No 1=Yes
6	Use gun	0=No 1=Yes
7	Jump from high places	0=No 1=Yes
8	Hanged himself or herself	0=No 1=Yes
9	Drown	0=No 1=Yes
10	Get hit by a car	0=No 1=Yes
11	Smoke/Gas	0=No 1=Yes
12	Others	0=No 1=Yes
Risk factors for self-harm		
No.	Attribute	Description
1	Resentment	0=No 1=Yes
2	Gossip	0=No 1=Yes
3	Disappointed love	0=No 1=Yes
4	Dispute with a close friend	0=No 1=Yes
5	Need care and support	0=No 1=Yes
6	Learning problems	0=No 1=Yes
7	Deceased parents	0=No 1=Yes
8	Family conflict	0=No 1=Yes
9	Chronic disease	0=No 1=Yes
10	HIV infection	0=No 1=Yes
11	Psychopathy	0=No 1=Yes
12	Depression problems	0=No 1=Yes
13	Elderly orphans	0=No 1=Yes
14	Addiction (individual)	0=No 1=Yes
15	Addiction (relatives)	0=No 1=Yes
16	Drinking alcohol	0=No 1=Yes
17	Economic problems	0=No 1=Yes
Health issues information		
No.	Attribute	Description
1	Diabetes	0=No 1=Yes
2	Hypertension	0=No 1=Yes
3	Heart disease	0=No 1=Yes
4	Chronic kidney failure	0=No 1=Yes
5	Chronic liver disease	0=No 1=Yes
6	Chronic lung disease	0=No 1=Yes
7	Chronic headache	0=No 1=Yes
8	Chronic pain	0=No 1=Yes
9	Neuralgia	0=No 1=Yes
10	Epilepsy seizure	0=No 1=Yes
11	Disabled	0=No 1=Yes
12	Paralysis	0=No 1=Yes
13	Cancer	0=No 1=Yes
14	AIDS	0=No 1=Yes
15	Psychosis	0=No 1=Yes
16	Others	0=No 1=Yes
Personal behavior		
No.	Attribute	Description
1	Smoking	0=No 1=Yes
2	Alcohol addiction	0=No 1=Yes
3	Addiction	0=No 1=Yes
4	Gambling addiction	0=No 1=Yes
5	Addicted to games/internet	0=No 1=Yes
6	Others	0=No 1=Yes
Death details		
No.	Attribute	Description
1	Suicidal ideation (class)	1=No 2=Yes

Next, consider records with errors included missing value, wrong input, values that enter above the actual range, such as age was 0 years or 100 years. Unreasonable data in each record was also removed. For example, suicide data stated that this

patient was dead, but suicidal ideation information of this patient is also available. From these steps, data set can be reduced up to 65,015 records, making the rest of the data is 129,474 records. After that, the data of the complete suicide was cut off amount 7,237 records. Finally, the rest of the data is 122,237 records.

Since the data in this research is mostly numerical data, the data must be converted to categorical data by using NumerictoNorminal function in Weka. The age is numerical already divided into five periods such as 10-27, 28-45, 46-62, 63-80 and 81-98 by using Discretize function in Weka to discretizes a range of numeric attributes.

C. Balancing the data with SMOTE

After the previous process, it was found that there was a large difference in data due to the fact that the data was not balanced. There are several ways of balancing the data, such as under-sampling and over-sampling. The under-sampling method is a technique used to reduce the amount of data which contained in a particular class. Therefore, it has to be close or equal amount of data in the fewer class by reducing the amount of data from more class. Whereas, the oversampling method works with fewer class by increasing amount of data in the fewer class to balance the data. The disadvantage of using this method is overfitting which improves the model to perform very well with training data. [5] We have tested the model with testing data and unseen data. The model can work, but it is inefficient. In this paper, the balancing the data with SMOTE is not reduced the amount of information and overfitting. Before the data is balanced, class "No" which means no suicidal ideation have 120,626 records. Class "Yes" which means suicidal ideation has 1,611 records. After balanced data by SMOTE, the total amount of data is 239,923 records as shown in Table 3.

Table 3. The total balancing data with SMOTE

Attribute	Value	Number of record before SMOTE	Number of record after SMOTE
Suicidal ideation.	No Yes	120,626 1,611	120,626 119,297
Total		122,237	239,923

Researchers tested on the dataset using 5, 7 and 10 fold cross-validation. The comparison result is shown in Table 4. The researchers found that 10 fold is the best result. We can increase the number of k-fold and make it more precise, but it takes longer to process, and the accuracy of the results was increased slightly. Usually only 10 fold cross-validation are used widespread. [6, 10]

D. Experiments with Weka

In the experiments, the Weka version 3.8.1 was used to evaluate the result of suicidal ideation. The 10 fold cross-validation, which is bold in Table 4, was used for evaluation because the results are the most accurate.

Table 4. Performance of K-Fold cross validation

Type	Method	Accuracy (%)		
		5 Fold	7 Fold	**10 Fold**
Single model	Naïve Bayes	70.38%	70.39%	**70.40%**
	ID3	90.55%	90.59%	**90.60%**
	C4.5	90.60%	90.63%	**90.69%**
Ensemble model	Random Forest	91.13%	91.14%	**91.17%**
	Bagging	90.79%	90.79%	**90.82%**
	Voting (C4.5, Naïve Bayes, Random Forest)	91.23%	91.25%	**91.26%**

E. Performance measurement

Cross-validation is a common technique for performance measurement of models because of its reliability. Cross-validation is the method which divides the data into several parts, usually represented by a k value, such as 5-fold cross-validation. First, the data is divided into five parts, each part has the same amount of data. After that, one of the data will be used as the performance tester and then it continues testing totally five times.

The accuracy can be calculated by equation (2). Usually, we used TP (True Positive), FP (False Positive), Precision, Recall and F-Measure to evaluate the classifier performance, which can be computed by equation (3) (4), and (5) respectively.

The accuracy (AC) is the ratio of the entire number of predictions that were correct. It is shown as the following equation:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2)$$

Precision (P) is the ratio of the predicted positive cases that was correct as estimated. It is shown as the following equation:

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

The recall or true positive rate (TP) is the ratio of positive events that were correctly identified as estimated. It is shown as the following equation:

$$Recall = \frac{TP}{TP+FN} \quad (4)$$

F-Measure is the average of precision and recall, as calculated. It is shown as the following equation:

$$F - Measure = \frac{2 \times precision \times Recall}{Precision+Recall} \quad (5)$$

IV. Experimental Result

For the SMOTE balanced dataset, it can be seen that in the single method, the most accurate model is C4.5, which gives the accuracy of 90.69%. In ensemble method, after combining the three classification schemes including C4.5, Naïve Bayes and Random forest, the researchers found that the best is ensemble voting, which gives the accuracy of 91.26% as shown in Table 5.

Table 5. Prediction Performance of Classifiers comparison

Type	Method	Accuracy (%)	Precision	Recall	F-Measure
Single Model	Naïve Bayes	70.40%	0.708	0.704	0.703
	ID3	90.60%	0.908	0.907	0.906
	C4.5	**90.69%**	**0.908**	**0.907**	**0.907**
Ensemble Model	Bagging	90.82%	0.910	0.908	0.908
	Random Forest	91.17%	0.913	0.913	0.912
	Voting (C4.5,Naïve Bayes, Random Forest)	**91.26%**	**0.914**	**0.913**	**0.913**

As the result, for a single model, C4.5 has an accuracy 90.69%, precision of 0.908, recall of 0.907 and F-measure of 0.907 which is higher than other methods. When using a tree structure to classify, we will be able to recognize the pattern of the person who has suicidal ideation as shown in the example. We are showing cases with a maximum accuracy of 2 cases.

Case 1: if "Psychopathy=yes and Age <=27
and Occupation=Labors and Alcohol addiction =yes"
then "Suicidal ideation = yes"

Case 2: if "Gambling=yes and Marital=Married
and Occupation= Agriculture and Gender=male"
then "Suicidal ideation = yes"

Case 1 means that if a person is laborer who has age not over 27, with psychopathy, and alcohol addiction, then this person has suicidal ideation. Likewise, Case 2 means that if a person is male, agriculture who get married and gambling addiction, then this person has suicidal ideation. We found the number of characteristics from C4.5 is 474 patterns. There are people, who have suicidal ideation over 500 people in each case is 42 patterns. There are 22 patterns that more than 90% accuracy. For ensemble model, voting method has an accuracy of 91.26%, precision of 0.914, recall of 0.913 and F-measure of 0.913 which is much higher than other ensemble classifiers.

V. CONCLUSION AND FUTURE WORK

Suicide is considered a problem for the country. Most suicides give some warning signs before committing suicide. If we are able to know the characteristics of the person who has suicidal ideation and are able to prevent from happening, it will reduce the number of deaths. This research aims to identify those at risk of suicide attempts by applying data mining techniques. This technique is the decision tree J48 (C4.5) compared with ID3 and Naïve Bayes. The C4.5 model yielded the best result with accuracy of 90.69%. After using decision trees, we could predict the characteristics of suicide attempters. In fact, the patient's data is not appropriate to synthesize. However, the decision tree cannot be utilized to find the characteristic because the number of data between suicidal ideation and no suicidal ideation is very different. It was necessary to balance data using SMOTE. We could use the ensemble method to get more accurate. When compared to other ensemble techniques namely bagging, voting and random forest, we found that ensemble voting method, which combines C4.5, Naïve Bayes and random forest, showed the best results in our study with 91.26% accuracy.

Future work is making the ensemble system more precise and accuracy as well as selecting the features to make the model faster. Therefore, it is necessary to develop the process of data management and select the appropriate techniques for the data. It would be beneficial if we could predict the time of committing suicide. We plan to build the application from this research that can help psychiatrist in preventing and supporting the surveillance suicide attempt.

ACKNOWLEDGMENT

We would like to thank the director and staff of Khon Kaen Rajanagarindra Psychiatric hospital for their help in collecting the data and their valuable technical support on this study.

REFERENCES

[1] Wan Sriruenthong, Wichira Pangchuntr, Kedsaraporn Kenbubpha and Supattra Sukawaha, "The Suicidality in Thai population:A National survey", in J Psychiata Assoc Thailand Vol.56 No.4, 2011

[2] Waraporn Prateepteranun, "The Rate of Suicide and Factors Related to Suicidal Behavior at Chophayayommarat Hospital", in THAIJO Vol.28 No.3 , 2014

[3] Thailand suicide rate report; Department of Mental Health, Ministry of Public Health : Thailand (2017)

[4] Han, J., Kamber, M., Classification. Data Mining: Concepts and Techniques. pp. 330-383. MA, USA: Morgan Kaufmann Publishers. (2012)

[5] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP: SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002, 16: 341-378.

[6] Wenjia Wang ,Graeme Richards and Sarah Rea , "Hybrid Data Mining Ensemble for Predicting Osteoporosis Risk", in Engineering in Medicine and Biology 27th Annual Conference, 2015 International Conference on, 2015, pp. :127-124IEEE.

[7] R. Quinlan (1986). Induction of decision trees. Machine Learning. 1(1):81-106.

[8] Ross Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA.

[9] M. N. Banu and B. Gomathy, "Disease forecasting system using data mining methods," in *Intelligent Computing Applications (ICICA),* 2014 *International Conference on*, 2014, pp. :133-130IEEE.

[10] Reda Al-Bahrani and Ankit Agrawal, "Colon cancer survival prediction using ensemble data mining on SEER data", in International Conference on Big Data , 2013 International Conference on, 2013 , pp. : IEEE

[11] Z.H. Zhou and Y.Jiang, "Medical diagnosis with c4.5 rule preceded by artificial neural network ensemble," in Transaction on Information Technology in Biomedicine 7, 2003 I, pp. 37-42 :IEEE.

[12] Shan Xu , Haoyue Shi ,Xiaohui Duan and Tiangang Zhu, "Cardiovascular Risk Prediction Method Based on Test Analysis and Data Mining Ensemble System," in Computing, Communication & Automation (ICCCA), 2015 International Conference on, 2015, pp. :127-124IEEE.

[13] Yan Li,Changxin Bai and Chandan K.Reddy, "A distributed ensemble approach for mining healthcare data under privacy constraints", in Information Sciences journal 330, 2016 I, pp. 245-259 :Elsevier.

[14] Neesha Jothi,Nur' Aini abdul Rashid and Wahidah Husain, "Data Mining in Healthcare", in The Third Information Systems International Conference, 2015 , pp. 306-313 :Elsevier.

[15] Eric Rojas,Jorge Munoz-Gama,Marcos Sepulveda and Daniel Capurro, "Process mining in healthcare", in Journal of Biomedical Informatics 61, 2016 , pp. 224-236 :Elsevier.

[16] Silviya D'monte ,Dakshata Panchal, "Data Mining Approach for Diagnose of Anxiety Disorder", in International Conference on Computing, Communication and Automation (ICCCA2015), 2015 :IEEE.

[17] Anjana Gosain,Amit Kumar, "Analysis of Health Care Data Using Different Data Mining Techniques", in IAMA 2009 , 2009 :IEEE.

Bayesian Network Optimization for Earthquake Prediction Using Simulated Annealing

Yosep Aditya Wicaksono[1], Defri Kurniawan[2], Heru Agus Santoso[3]
Computer Sciences Faculty of Dian Nuswantoro University
of Dian Nuswantoro University
Imam Bonjol Street 207, Semarang
Correspond: [1]yosepaw94gaul@gmail.com, *[2]defri.kurniawan@dsn.dinus.ac.id, [3]heru.agus.santoso@dsn.dinus.ac.id*

***Abstract*— Earthquakes are natural disaster that can be very harmful. In case of earthquake, the data are recorded by the seismograph station. With the data, it will be carried out the prediction process using Bayesian Network, to obtain a technique which has the best accuracy in the earthquake prediction. This study compare the performance prediction between Bayesian Network alone with Bayesian Network which is optimized using simulated annealing. From the results of research conducted is proven that Simulated Annealing can improve the accuracy of earthquake prediction on BMKG online data by increasing 2.23% of accuracy**

Keywords: Naive Bayes, Bayesian Network, Simulated Annealing, Optimization

I. INTRODUCTION

Earthquake is a kind of disaster which can be dangerous if it happens on large scale. An example of a major earthquake happened in Indonesia is the earthquake in Aceh, on Sunday, December 26, 2004[11]. There were many victims at that time because of unreadyness. They didn't have enough time to escape and avoid the disaster. Therefore, early warning system which is carried out by earthquake prediction is needed. However, the current approaches mostly use time series prediction and some of them use multiple variables as input for prediction which do not take into account the relation among the variables. In other words, this kind of approaches can be considered ignoring the fact that involved variables actually have relationship to each other.

As a matter of fact, every earthquake has interrelated variables. Hence, in our approach, the prediction of earthquake takes advantage of the relations among variables. The Bayesian network is then used as chosen approach. The network means to handle the related variables. However, in the previous research the accuracy still needs to be improved, to produce the best optimization solution approach. it is necessary to predict earthquakes to improve the accuracy of prediction performance using Bayesian Network with Simulated Annealing [3]. In this study, simulated annealing optimizes global optimum solution for the problem prediction. With this optimization technique, the accuracy of earthquake prediction when handling tectonic earthquake prediction can produce the most optimal value of Bayesian Network.

As mentioned earlier, the aim of this study is to improve earthquake prediction accuracy of Bayesian Network which is carried out by means of Simulated Annealing. The increasing of the accuracy with produce a better prediction approach, so that people will be more trust in response to disaster. It is important because the damage to the building will be very dangerous with potentially incidence of casualties. The more responsive to the disaster, the less to loss life of victims. And for science and technology contribution, this research can be a source of references for further development. Especially in the case of earthquakes, it is still quite rare that examines the earthquakes prediction are computed with high accuracy which are available in the references.

II. RELATED WORK

Earthquakes have interrelated involved variables and they affect each other during earthquakes. Bayesian Network aims to handle these influencing variables [1]. Earthquakes can not only destroy building but also as trigger to landslides. They also occur because of influential variables in the event. Bayesian Network was used to deal with influencing variables. However, it still did not have good accuracy in handling large data and occurs continuously [2]. In the next work, Bayesian Network would be formed when looking for probabilities of Most Probable Explanations (MPE). MPE is produced using genetic simulated annealing. The result of this study is promising[3]. Our preliminary study of earthquake detection using mini detector is also presented in [4]. However, all of those works will be more complex in handling large numbers of variables. it is necessary to optimize our approaches using simulated annealing where the network needs to be reconstructed based on the initial networks to produce the best accuracy.

From the previous research, it can be said that Bayesian Network is promising, as it is very reliable in handling large and complex probability data. But it cannot provide good accuracy when handling large and continuous data [1]. There is a study which stated that Bayesian Network uses simulated annealing can provide a better probability solution when handling large amounts of data.

III. METHODOLOGY

A. Data Collection

Data used in this research is data taken from public data of earthquake per day at the official site of BMKG Indonesia http://dataonline.bmkg.go.id. The amount of data are 403 data with a span of time 01-07-2017 up to 25 -07-2017. The attributes are latitude, longitude, depth, mag, rms.

With the data, it will be carried out the optimization process using Bayesian Network, to obtain a technique which has the best accuracy in the earthquake prediction. Especially if the data used is a large scale and continuous data.

B. Pre-Processing Data

Out of 403 data will be taken 90% for training set and 10% for testing set. Therefore, for the next study, we will add more data which could improve the prediction accuracy. We consider this works as initial work for further improvement. The preprocessing stage consist of the following steps:

- Data Editing

This step is performed to fix incomplete data which obtained from BMKG. There are some data that have particular attributes need to be filled. Then the value of this kind of attributes are filled with the means value their neighborhood.

- Data Discretization

This step is to make the data discrete, which is to make numerical data to be categorical data. It is very important to do discretization for numeric data because this data will be processed using Bayesian Network with the basis of probability calculation.

C. Modeling

After getting the data that has been preprocessed, the next step is to model the selected attributes with the following process:

- Forming Networks

In the beginning, we firstly form a Naïve Bayes model to identify involved initial set of variables. The result of the model will be compared with Bayesian Network optimized with Simulated Annealing. The model is formed based on causal parameters that occur based on attributes data identification taken from BMKG data and form the network as below.

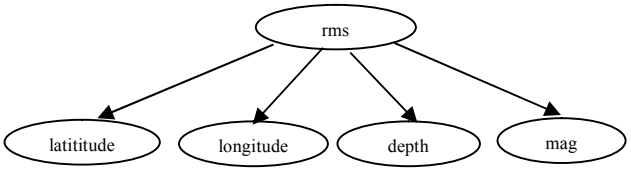


Fig. 1. Initial Bayesian Network

The next step is Bayesian Network optimization with Simulated Annealing

Simulated Annealing plays important role in establishing the initial Bayesian Network, and it will select the variables that have the closest interconnection by identifying each variable, which will result in a more optimal new network form.

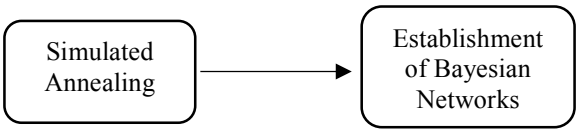


Fig. 2. Scheme of Bayesian Network Optimization with Simulated Annealing

- Bayesian Network calculations

a. Network before optimized with Simulated Annealing

The process is started with the propability calculation of continuous data of earthquake, i.e., $P(X_i = x_i \mid Y_j)$ for each attribute which are rms, latitude, longitude, depth, mag.

With each attribute of rms, latitude, longitude, depth, mag compute the probability of earthquake with the following formula:

$$P(X_i = x_i | Y_j = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}}} e^{\frac{(x_i - u_{ij})^2}{2\sigma^2 ij}} \quad (1)$$

Note:

P: Probability of earthquake

X_i: attributes of earthquakes (rms, latitude, longitude, depth, mag)

x_i: The value of earthquake attributes (rms, latitude, longitude, depth, mag)

Y_j : the sought after earthquake Class (N, S, W, E, In, Medium, Shallow, Big, Medium, Small)

y_j: a subclass Y_i to search (N, S, W, E, Deep, Medium, Shallow, large , Small)

u : Mean, expresses the average of all attributes

σ is the Standard Deviation, declares a variant of all attributes

b. Optimization with Simulated Annealing

Simulated Annealing (SA) is used to form Bayesian Network nodes more effectively and means to produce good accuracy. Here is the formula of Simulated Annealing

$$p = e^{\frac{-\Delta E_i}{T(t,i)}} \quad (2)$$

SA works based on the analogy that in the beginning of SA, it is chosen the initial solution which represent the initial material condition before the process has begun. Free movement of atoms in material is represented in form of early solution. This solution can be modified freely.

TABLE I. Naïve Bayes Node

State	Node
State 1	T(rms,latitude)
State 2	T(rms,longitude)
State 3	T(rms,depth)
State 4	T(rms,mag)

P is the probability, ΔE_i is the state formed from the node as in table 1 is state 1, state 2, state 3, and state 4. T ((i, j)) is the node, which will be T (rms, latitude), T (rms, longitude), T (rms, depth), T (rms, mag).

IV. EXPERIMENTS AND RESULTS

In this research, it is found that Bayesian Network which is optimized using SA. SA is a very important role in forming Bayesian Network. As a result, the network of variable formation is created as follows.

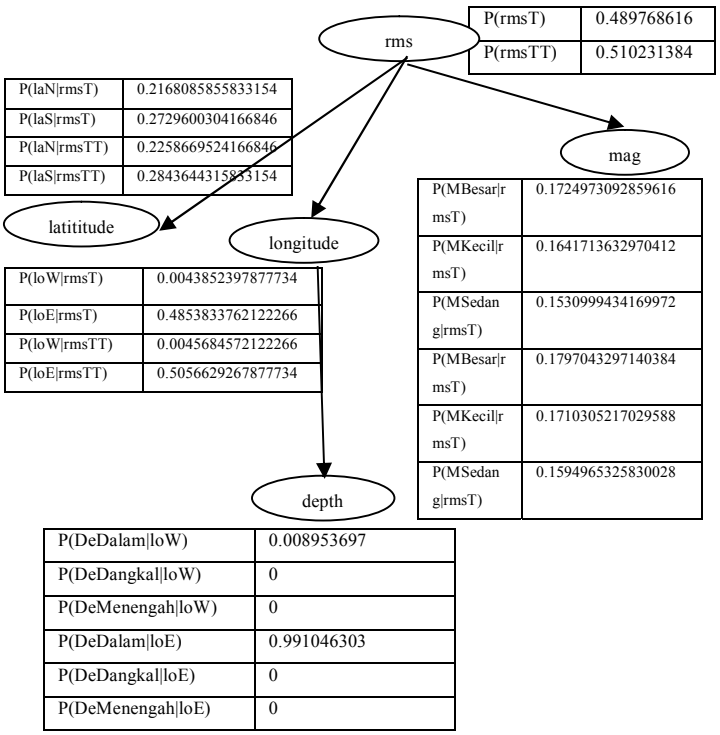


Fig. 3. Results Of Optimization Using Simulated Annealing

After that, the optimization performance measurement performed by Simulated Annealing is obtained with the following measurement results:

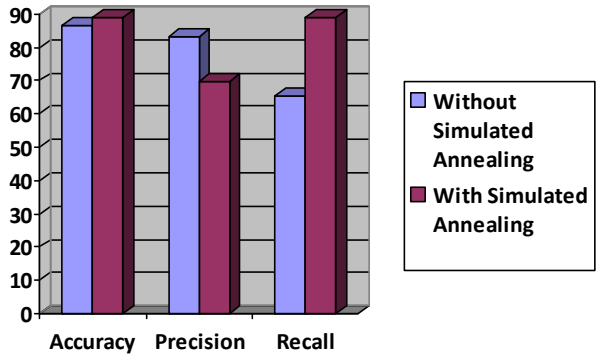


Fig. 4. Bayesian Network Comparison Results

From these results obtained, Bayesian Network optimized using Simulated Annealing has gained an increasing Accuracy as of 2.23%.

V. CONCLUSION

From the results of research conducted is proven that Simulated Annealing can improve the accuracy of earthquake prediction on BMKG online data. By using Bayesian Network without using Simulated Annealing, the result reaches of 86.6004% of accuracy. On the other hand, while using Simulated Annealing, it reaches 88.8337% of accuracy, it has better accuracy than previous research[1] only has 85% of accuracy. Hence, from the result shown, there is an increasing 2.23% of accuracy. In addition, these results can also be said that the earthquake that occurred in Indonesia occurred in shallow and medium depths.

REFERENCES

[1] Xiao Fan, Shoudong Han, Yong Zhao, "An Approach Bayesian N(Wikipedia, 2018)etworks In Magnitude Forecast Based On Earthquake Trace Cloud," in *IET International Conference on Information Science and Control Engineering*, China, 2012.

[2] Yiquan Song, Jianhua Gong, Sheng Gao, Dongchuan Wang, Tiejun Cui, Yi Li, Baoquan Wei, "Susceptibility assessment of earthquake-induced landslides using Bayesian network: A case study in Beichuan, China," *Computers & Geosciences,* vol. 42, p. 189–199, 2012.

[3] A. M. Abdelbar and S. M. Hedetniemi, "A parallel hybrid genetic algorithm simulated annealing approach to finding most probable explanations on Bayesian belief networks," in *Neural Networks,1997., International Conference on*, USA, 1997.

[4] Y. A. Wicaksono and H. A. Santoso, "Implementasi Metode Bayesian Network Untuk Decision Support System Pada Mini Detector Earthquake," in *Seminar Nasional Teknologi Informasi dan Multimedia 2016*, Yogyakarta, 2016.

[5] D. Widiastuti, "Analisa Perbandingan Algoritma Svm,

Naive Bayes, Dan Decision Tree Dalam Mengklasifikasikan Serangan (Attacks) Pada Sistem Pendeteksi Intrusi," in *Universitas Gunadarma*, Depok, 2007.

[6] A. I. Pristanto, "Upaya Peningkatan Pemahaman Masyarakat Tentang Mitigasi Bencana Gempa Bumi Di Desa Tirtomartani Kecamatan Kalasan Kabupaten Sleman Provinsi Daerah Istimewa Yogyakarta," Universitas Negeri Yogyakarta, Yogyakarta, 2010.

[7] A. A and R. G, "Earthquake Early Warning System by IOT using Wireless Sensor Networks," in *IEEE WiSPNET 2016 conference*, Salem, 2016.

[8] Y. A. Wicaksono, N. R. Susanto and S. Fahriah, "Implementasi Metode Naive Bayes Untuk Detektor Gempa Sederhana Yang Tersinkronasi Dengan Handphone," in *Digital Information and System Conference Universitas Kristen Maranatha*, Bandung, 2015.

[9] BMKG, "Skala MMI (Modified Mercalli Intensity)," BMKG, 22 11 2016. [Online]. Available: http://www.bmkg.go.id/gempabumi/skala-mmi.bmkg. [Accessed 22 11 2016].

[10] M. A. Ashraf and M. H. Sandra, "A Parallel Hybrid Genetic Algorithm Simulated Annealing Approach to Finding Most Probable Explanations on Bayesian Belief Networks," in *International Conference on Neural Networks,1997*, Houston, 1997.

[11] Wikipedia, "2004 Indian Ocean earthquake and tsunami," Wikipedia, 25 08 2018. [Online]. Available: https://en.wikipedia.org. [Accessed 26 08 2018].

Automatic Gestational Age Estimation by Femur Length using Integral Projection from Fetal Ultrasonography

Khusnul Danny Rahayu [1)], Riyanto Sigit [2)], Dias Agata [3)], Agung Pambudi [4)], Nurul Istiqomah [5)]
1,2,3,4,5) Department of Informatics and Computer Engineering
Electronics Engineering Polytechnic Institute of Surabaya
Surabaya-East Java, Indonesia
{[1]khusnuldrahayu, [4]agungpambudi, [5]nurulistiqomah}@ce.student.pens.ac.id, {[2]riyanto, [3]diasagata}@pens.ac.id

***Abstract*—The accuracy of scanning is important to properly know a fetus condition in the womb. However, the results still depend on a doctor's observation, so there are possibility of errors in identification. We propose an automatic gestational age identification system of images using femur length parameters to reduce identification errors. The method used is a combination of Gaussian Filter, Morphology Operation, and Canny Edge Detection methods to improve scanning result and to minimalize the noise. Find Contour method is used to obtain the fetal femur object, then searched for femur length using Integral Projection. This study uses samples of fetal femur data in second trimester. From the result of the study Integral Projection can be used to measure the femur length with an average success rate of 94.8% and the identification of gestational age with an average success rate of 97.6%.**

Keywords—Fetus; Automatic; Gestational Age; Femur Length; Integral Projection;

I. INTRODUCTION

There are several factors that can help the success in pregnancy, which are by calculating the age of pregnancy, identifying the development of the womb, and knowing the estimated delivery due time. By knowing the age of pregnancy, a mother can know the development or the growth of the organ in the fetus, what the fetus needs and what things are allowed and prohibited to do during the gestational age. In addition, by knowing the age of pregnancy, a pregnant woman can estimate the examination schedule that must be done with the doctor or the midwife, so that the pregnancy will not face any problem and will lead into a successful delivery [1].

Gestational age can be known by many ways, one of them is through ultrasonography (USG). To know the age of pregnancy through USG, there are several parameters used, one of them is the measurement of fetus femur. But, it is alone may not necessarily produce accurate information because the resulted image has a low quality.

The results also depend heavily on the observation of the physician or the operator using it, so there is no possibility of errors in identification [2]. To reduce the error of identification, we create a system that can detect the age of pregnancy automatically using fetal femur.

The method we use is a combination of Gaussian Filter method, Morphology Operation, and Canny Edge Detection to improve the image. Find Contour used to get the fetal femur object. And Integral projection is used to obtain fetal femur length. This study also aims to make it easier for doctors to get more accurate gestational age from the fetal femur through ultrasound. With this system the doctor no longer needs to pull certain points to get the fetal femur length. The system will automatically get objects along with the size of the fetal femur length.

II. RELATED WORK

Previous researches also discussed the measurement of fetal biometry through the length of the fetus, and the fetal head. The length of the fetus is measured using Crown Rump Length (CRL) [3], [4] and the fetal head is measured using Biparietal Diameter (BPD) [5], [6].

Gusti Ayu Almira and Ahmad Puji Santoso has built a system to analysis of the fetal health condition. The research using Crown Rump Length (CRL) and the gestational sac diameter to get the basis for determining the level of the fetal health. Gestational age from the CRL and gestational sac would be compared with the calculation of the first day of last menstruation (HPHT). This research has an error rate of 5%. CRL is preferred to the measurement of gestational age at the beginning of pregnancy or first trimester.

Whereas in the study using Biparietal Diameter from Lokesh Kumar has built a system to compare the accuracy of predicting the gestational age using Biparietal's Diameter measurement in second and third trimester. Prediction of gestational age before 30 weeks has an accurate result but the precision declines after that. Measurement using Biparietal Diameter often having trouble when engaged fetal head, direct occipito-anterior and occipito-posterior positions and in breech presentation. The results of the study have an accuracy rate of 66.28%.

III. METHOD AND MATERIAL

The femur picture of the fetus is obtained from the measurement of a specialist gynecologist who manually that is, by capturing an image that is good if possible, then pulling the point from one part to another on the fetal femur using a scroll mouse on an ultrasound machine. Images taken from patients with 16 to 23 weeks gestation (second trimester).

Machine used to take pictures of fetus femur is Alpinion E-CUBE 9 Diamond with jpg format. The flow chart system for gestational age estimation of the femur length can be described in Fig. 1.

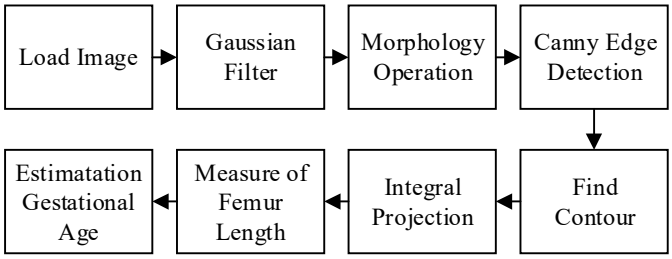


Fig. 1. Flowchart system

A. Gaussian Filter

Gaussian filter efficient to smoothing images and used to remove noise and detail [7]. This is as a result of excessive noise intensity. How Gaussian works is to remove the high-frequency components of the image.

So the Gaussian Technique is said to be a low-pass filter method. The Gaussian filter is obtained by multiplying the pixel value of the original image and the kernel in the form of a 2D Gaussian function.

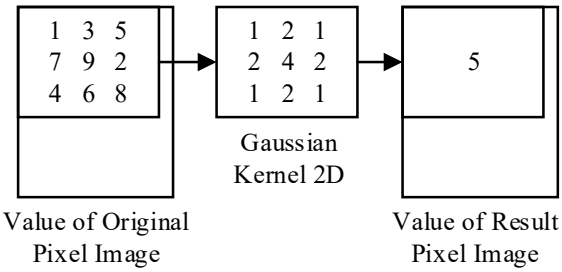


Fig. 2. Gaussian 2D function

Gaussian filter uses Gaussian distribution function. Then the filter is generally represented in the form of a two-dimensional array on $[x, y]$. Gaussian equations for two-dimensional space.

$$G(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}} \quad (1)$$

$G(x,y)$ is a gaussian matrix element at position $[x, y]$ is the standard deviation or sigma. The bigger: localization (distance between pixels) is weak or remote, but to detect (edges, noise, etc.) gets better. The smaller: localization (distance between pixels) is good, but for detection (edges, noise, etc.) is weak. is the size of the Gaussian matrix that reaches the point $-x$ to $+x$, and its midpoint is at $x = 0, y = 0$.

B. Morphology Operation

The morphological operation are used to detect noisy pixels in the residue detector. Since the open removes salt noise and close removes pepper noise, through the operators the salt and pepper pixels noises can be notably determined. In general, these transformations find structures which have been removed by the opening and closing filters and the residual between the original image and the filtered image increases notably the contrast of the erased regions [8].

In this research the morphological operation used is Top-hat. Top-hat results from combining image reduction with opening and closing. The main principle of this transformation is to remove objects from images by using Structuring Elements in opening and closing operations that do not match the objects to be deleted. Top-hat transforms are used for bright objects in dark backgrounds, The equations of top-hat transformations are as follows:

$$T_{hat}(f) = f = (f \; o \; b) \quad (2)$$

The top-hat transform of the grayscale image f is defined as f minus the opening result.

C. Canny Edge Detection

In the process of Canny edge detection, the edges of the image objects will be generated, which aims to mark the part of the image detail and improve the detail of the image is blurred, which occurred due to error or the effect of image acquisition process.

The steps to perform the Canny Edge Detection [9] process are as follows:

1. First, smoothen the image with image convolution and Gaussian filter g (x, y) For digital image processing, zero mean Gaussian two variables are expressed by the following equation:

$$\frac{G(x,y)}{c} = e^{-\frac{x^2+y^2}{2a^2}} \quad (3)$$

2. Second, find the gradients on each pixel by applying a sobel operator. The equation for estimating the gradient in the x and y directions by using the kernel as follows:

$$K\,Gx = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad (4)$$

$$K\,Gy = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \quad (5)$$

3. Next, we must determine the edge direction to improve the quality of the image, with the equation as follows:

$$\theta = arctan\left(\frac{|Gy|}{|Gx|}\right) \quad (6)$$

θ is the angle of gradient used to determine the direction of pixel comparison in the next process.

4. The next step is non-maximum suppression or peak detection which is used to maintain its gradient points which have the highest intensity value of its neighbors. This process is done by utilizing direction information from an edge.

5. Then, the thresholding process is done to check the pixel value against the threshold value, if the pixel value is

greater than the upper threshold, then the pixel is recognized as the strong edge, with the equation as follows:

$$T = {(f_{max} + f_{min})}/{2} \quad (7)$$

If the pixel value lies below the upper threshold and below the upper threshold then the pixel is recognized as a weak edge.

6. Finally, the point recognized as the strong edge and the weak edge obtained from double thresholding will be checked. If the pixel value is greater than the upper threshold then the pixel is the edge pixel. If the pixel is smaller than the lower threshold then the pixel is ignored.

D. Find Contour

Find contour here is used to remove the small noise that still exists after the canny edge detection process. Here are the stages of the Find Contour process.

1. Obtain the input image which will be used in find contour processing.
2. Change the input image to grayscale.
3. Blur the image.
4. Perform Edge detection using Canny Edge Detection.
5. Perform Edge detection results are used to search for contour.
6. Calculate the area of each contour that has been found and gives a limit if > 300 then contour is drawn.
7. Draw contour that has an area of > 300.

E. Integral Projection

An integral projection is a method used to locate the region or location of an object. This method can be used to detect the boundaries of different image areas, so it can search the location of faces and features. This method can also be called an integral row and a pixel column, since this integral adds pixels per row and pixels per column [10]. The following is the integral projection equation used.

$$HIP = \sum_{j}^{n} 1f(i,j) \quad (8)$$

$$VIP = \sum_{i}^{n} 1f(i,j) \quad (9)$$

The picture below is an illustration of the integral projection method.

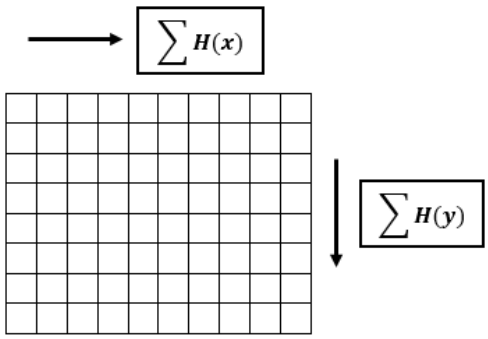


Fig. 3. Illustration of integral projection method

In this study vertical projection is used to find the starting point and end point of the femur segmentation of the fetus. The image used in this method is grayscale type.

F. Femur Length

Femur Length measurements were obtained from measuring the tip of the femur to the other end of the femur which has been obtained from the integral projection method using the Euclidian Distance formula. Here is the Euclidean distance formula.

$$Euclidian\ Distance = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (10)$$

After results obtained from the Euclidean distance which is the length of the femur it can be searched gestational age by comparing the femur length results with standardized table Hadlock [11]. The standardization table is in the Table 1.

TABLE I. STANDARIZATION TABLE

GA (Week)	FL (cm)	GA (Week)	FL (cm)
13	1.07	27	5.14
14	1.40	28	5.38
15	1.73	29	5.61
16	2.05	30	5.84
17	2.37	31	6.06
18	2.67	32	6.27
19	2.98	33	6.48
20	3.27	34	6.68
21	3.56	35	6.87
22	3.84	36	7.06
23	4.11	37	7.23
24	4.38	38	7.41
25	4.64	39	7.57
26	4.89	40	7.73

IV. RESULT

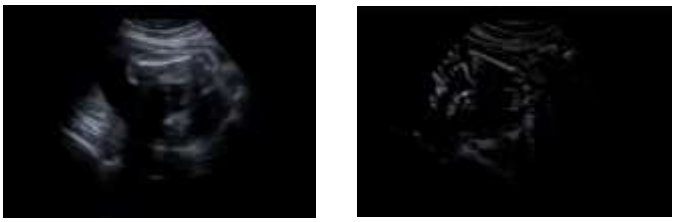
The first stage in this research is loading the image of fetus femur. The fetus femur image can be seen in Fig, 4.



Fig. 4. Result of fetal femur

Images have a high enough noise level so that the preprocessing process needs to be done so as to reduce the noise contained in the images of fetus femur.

The preprocessing process used in this research is Gaussian Filter, Morphology Operation, Canny Edge Detection. The results of the preprocessing process can be seen in Fig. 5.



a. Gaussian filter b. Morphology operation

c. Canny Edge Detection d. Find contour

Fig. 5. Preprocessing process result

The results of the Gaussian Filter process have a blur effect instead of the previous image. The Gaussian Filter results are in Fig. 5 (a). Then the next process is Morphology Operation, the result is a picture of femur has a good illumination level when compared to previous images. To get the edge of the object then used Canny Edge Detection.

From the process of Canny Edge Detection still looks some noise. Noise in the intent is the small parts that exist between the objects of the femur. To fix the shortcomings of the Canny process then use Find Contour to select small parts. So, the object will look more clean. The find contour results are shown in Fig. 5 (d).

From the results of find contour process sought the starting and ending points of the femur using integral projection. Each pixel in the horizonal region is scanned from the left to find the starting point and then after finding the pixel with an intensity > 100 then the pixel is again scanned from the right to find the end point with an intensity pixel > 100.

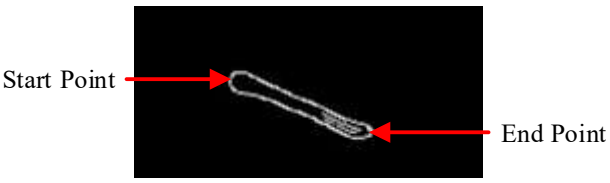


Fig. 6. Integral projection

Having obtained the starting and ending points of the femur image then to find the length of the femur used Differential Euclidian formula. The result of Euclidian distance is shown in Fig. 7.



Fig. 7. Femur length

After results obtained Euclidean distance it can be known the length of the femur and can be sought age of pregnancy. From several experiments yield different levels of accuracy. The results and percentage errors are listed in Table I.

TABLE II. FEMUR LENGTH MEASUREMENT

File Name	System Result (cm)	Manual Result (cm)	Percentage (%)
1.jpg	4.38	4.18	95.2
2.jpg	4.8	4.63	96.3
3.jpg	4.26	4.38	97.3
4.jpg	1.8	1.99	90.5
5.jpg	3.84	4.14	92.8
6.jpg	3.72	3.96	94
7.jpg	2.1	2.05	97.6
Error Average			94.8

Table II is a table of femur length measurement results of the system that has been made and from the doctor's manual along with the percentage of success of the system that has been made. The maximum percentage of success rate is 97.6% of the seven data that have been tested and the average success rate is 94.8%.

TABLE III. GESTATIONAL AGE RESULT

File Name	System Result (cm)	Manual Result (cm)	Percentage (%)
1.jpg	24	23	95.7
2.jpg	25	25	96
3.jpg	24	24	97.9
4.jpg	15	15	100
5.jpg	22	23	95.7
6.jpg	21	22	97.7
7.jpg	16	16	100
Error Average			97.6

Table III is a table of gestational age results of the system that has been made and from the doctor's manual along with the percentage of success of the system that has been made. The maximum percentage of success rate is 100% of the seven data that has been tested with an average success rate is 97.6%.

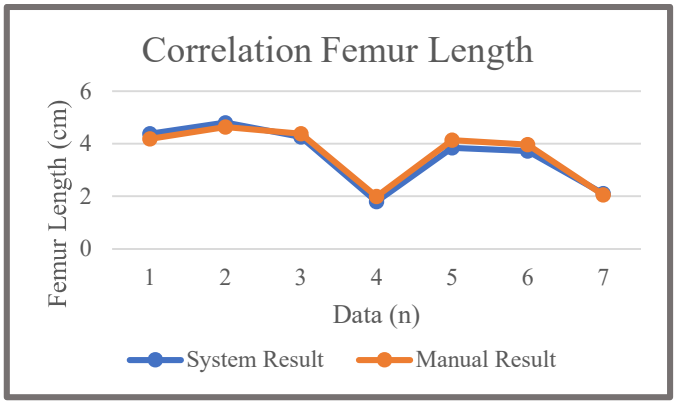


Fig. 8. Femur length correlation

Fig. 8 is a correlation graph of femur length between systems that have been made with doctors' manual results. The results of the system and also the results of doctors have a fairly high degree of similarity there are only slight differences and can still be categorized successfully because it has no significant difference.

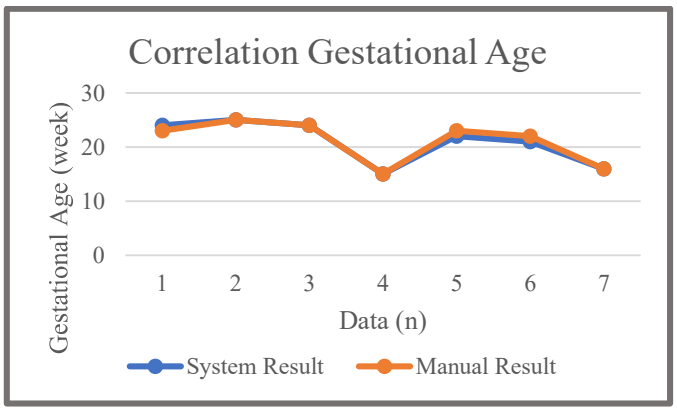


Fig. 9. Gestational age correlation

Fig. 9 shows a correlation graph of pregnancy age between systems that have been made with the results of the physician manual. Results of correlation of gestational age more have a higher degree of similarity when compared with the results of femoral length correlation. That's because the difference in the length of the femur is still within the same age range so it still has the same age value although the length of the different femur.

V. Conclution

The proposed automated method has a strong correlation with the manual system, therefore the measurement result between the proposed system and the manual system used by the doctor is quite similar. The average success rate of femur length measurement is 94.8% and the average success rate of gestational age measurement is 97.6%. Consequently, we can conclude that the proposed method is useful to improve clinical diagnosis result by minimalizing the error.

Acknowledgment

We thank Obstetric and Gynecology Doctors of Graha Masyithoh Surabaya who had contributed to this research by providing the data of fetus femur. We also thank to the Ministry of Research, Technology, Higher Education of Indonesia who had funded this research.

References

[1] L. N. Sufrastya, “Perancangan Sistem Informasi Kandungan Ibu Hamil”, Ponorogo : Universitas Muhammadiyah Ponorogo, 2015.

[2] P. Adhi, M. C. Johanes, and W. F. Firman, “Ultrasonografi Obstetri & Ginekologi.” Jakarta: C.V Sagung Seto, 2011.

[3] Gusti Ayu Almira, Tri Harsono, Riyanto Sigit, I.G.N.T.B.Bimantara, Michael Saputra J.S., “Performance Analysis of Gaussian and Bilateral Filter in Case of Determination the Fetal Length”, IEEE KCIC, 2016.

[4] Ahmad Puji Santoso, Riyanto Sigit, “Health Monitoring of Fetal Ultrasound Image Using Active Contour Models”, IEEE iSemantic, 2017.

[5] Naiad Hossain Khan, Eva Tegnander, Johan Morten Dreier, Sturla Eik-Nes, Hans Torp, Gabriel Kiss, “Automatic Measurement of Biparietal Diameter with a Portable Ultrasound Device”, IEEE International Ultrasonics Symposium Proceedings, Chicago, IL, USA, 2014.

[6] Lokesh Kumar, Bhawani S. Modi, Shikha Sharma. “Ultrasonographic Fetal Gestational Age Determination by Biparietal Diameter”. Annals of International Medical and Dental Research (AIMDR) Assistant Professor, Dept. of Radiodiagnosis, FHMC, Tundla, UP. Dept. of Anatomy, FHMC, Tundla, UP. 2017.

[7] J. J. Hwang, K. Hyeon Rhee, “Gaussian Filering Detection based on Features of Residuals in Image Forensics”, IEEE RIVF International. Hanoi-Vietnam, 2016.

[8] Anjanappa. C., Sheshadri. H. S, “Development of Mathematical Morphology Filter for Medical Image Impulse Noise Removal”, IEEE ICERECT International Conference on. Mandya, India, 2015.

[9] Abdiansah, Rozki Romodhon, “Ekstrraksi Bentuk Janin pada Citra Hasil USG 3 Dimensi menggunakan Deteksi Tepi Canny”, JRSCA, Vol. I, No. I, 2012.

[10] Chen Dong, Chen Pei-hua, Xiaoli Wang, Yang Pu-liang, “Eye Detection Based on Integral Projection and Hough Round Transform”, IEEE, 2015.

[11] SuperSonic Imagine, “Obstetrical References”, Cedex : France, 2016.

A Lexicon-based Sentiment Analysis for Amazon Web Review

[1]Ardianda Aryo Prakoso, [2]By Winantesa Yananta, [3]Arif Fitra Setyawan, [4]Muljono
[1,2,3,4]Department of Informatics Engineering
Dian Nuswantoro University
Semarang Indonesia
[1]ardianda.prakoso@gmail.com, [2]winantesa25@gmail.com, [3]arifprasetya@gmail.com, [4]muljono@dsn.dinus.ac.id

***Abstract*-The development of internet more quickly over time as well as the development of e-commerce one of them is amazon.com. The amazon.com is one of the largest e-commerce in the world by providing various needs that can be accessed by internet. Review feature provided to make amazon.com party can know the various responses from consumers. However amazon.com difficulties in summarizing the various kinds of reviews that are positive or negative. By using one natural language processing the main purpose is to help the amazon.com in knowing the most responses from consumers to improve the quality service. In this research we will be using the dataset that was obtained from UCI Machine Learning that contain 1000 set of data which has 478 of negative data and 522 positive data and this will be combine a variety of classification methods for comparison and in preprocessing process will be added with lexicon technique to improve the quality of preprocessing. The result of this research is K-Nearest Neighbor with lexicon technique get highest accuracy with value 92.67% followed by SVM with lexicon get 91.33% accuracy and last Decision tree with 82% accuracy.**

Keywords—Sentiment Analysis, amazon.com, classification, lexicon

I. INTRODUCTION

The rapid development of technology creates ease in accessing a variety of things one of them find items that we want on the internet. The development of conventional markets to e-commerce is increasing as the rapidly evolving technology develops. The customer that using e-commerce service continue to increase from 2013 by 1.3 billion customer to 1.7 billion by 2017 for the rest of the world [1]. One of the largest e-commerce in the world is amazon.com [2]sell a variety of items such as e-books, software, CD / DVD, and many more [3].Being one of the largest retail stores amazon.com always improve its service one of them by applying a review system to the product where the review is done by consumers who have purchased products on amazon.com. There are various reviews given that have positive or negative traits.

To improve the quality of service, amazon.com sees rating of its products based on positive or negative responses given by customers. But amazon.com has many products and customers so it is difficult to see the reviews one by one so that is required a method that can be applied in assessing various kinds of responses automatically the example one of the methods of sentiment analysis.

Sentiment analysis or also called opinion mining is one of natural language to do word processing that can track public opinion about a certain product or topic [4].Sentiment analysis will divide the words, then look for the equation and make it one that will later be divided into positive, negative, and neutral [5].

One method of sentiment analysis for text data is using text mining techniques. Basically text mining is a technique for classifying documents where this technique will try to find the unique pattern of a large set of text on the data [6]. Text mining consists of four stages. The first step is Information Retrieval the step that involves to find out the collection of text which might need some conversion, the second step is Natural Language Processing which analyze the text in structure, the third step is Information Extraction this step main purpose is to find out the relationships between the texts, and the fourth step is Data Mining this step main purpose is to find any similarities between the information that has the same meaning [7].

In this study we will compare the classification method to find the best method in text mining such as Decision Tree, SVM, and K-Nearest Neighbor and also will be compared with the integration of each method with Lexicon. The dataset used is the amazon.com dataset review obtained from UCI Machine Learning which contains about positive and negative reviews from amazon.com consumers.

II. RELATED WORK

Research using Sentiment Analysis has been widely used primarily in finding the best accuracy in every problem encountered in large sites like social media or large e-commerce to assess the various responses provided by the community. Here is a related study that can help in this research.

The research conducted by Rastogi et al [8]developing lexicon sentiment analysis combined with the SVM classification method. Based on the research, the sentiment analysis approach using machine learning is better than the lexicon based approach, although machine learning approach ignores the knowledge that lexicon method has made in word processing. In this research sentiment analysis based on lexicon will be the main information which will be processed by SVM method as machine learning. The dataset used is a dataset created by Blitzer et al where the dataset is used for research purposes. The results obtained from this study prove an increase

in accuracy on the results. Using SVM method the accuracy is 42.3%, while combining lexicon domain method, MPQA, and SVM is 49.3% which is 7% increase in the performance conducted by this research.

According from research by Olga et al [9] the develop of lexicon method by combining it with machine learning method to improve accuracy where the case in this research is sentiment analysis from Twitter. Combining the lexicon of emotion, lexicon, N-gram and combining the cost sensitive method with SVM yields accuracy for 64% Naïve Bayes, 62% Decision Tree, SVM 66%, and Cost Sensitive SVM method obtains the highest accuracy result with a yield of 73%.

The Lexicon can be used for variety of language depending on how it use may be variety just such as these research by Lucas V.Avanco [10] The data was gathered from one of the traditional online service in Brazil called Buscape, it contain 2000 review which are 1000 were set as positive and 1000 were set as negative these classification is based on the review final recommendation about the product in Brazilian Portuguese language by using 3 sentiment lexicon-based extension that specialize for Brazilian Portugese language which are OpinioLexicon, SentiLex and LIWC, the result show that the better result show that the sentiment SentiLex got the best result achieving 74% of accuracy while the UnionLexicon had the worse result, the possible reason for that it's had problem of reading the sentiment word that had dependent on the context such as "grande (big)" and "pequeno (small)".

Twitter can become one of the source that has plentiful of data for the sentiment analysis and opinion mining, from the Alfan [11] the research purpose is to automatically building a training corpus for the sentiment analysis on Indonesian tweets, the data was gathered from May 16th 2013 to June 26th 2013, by using Twitter Streaming API and geo-location to filter tweets posted in the area of Indonesia, the result are 637 that came from different various of domain these tweet has been tagged manually by the human annotators. The evaluation was compared using Naïve Bayes and Maximum Entropy as the evaluation model with 3 pre-processing method the baseline method, Lexicon based method and Clustering based. It is conclude that the lexicon based has the best result with 81.13% accuracy for Naïve Bayes and 86.82% for Maximum Entropy and it outperform the Baseline and Clustering method. These research was based on Vania [12] which has proposed method in domain-specific sentiment lexicon for Indonesian language the main purpose is to expand the usage of using sentiment patterns and to classify polarity of a word using sentiment score. The method is able to generate sentiment lexicon automatically by using a small seed of sentiment words, part of speech (POS) tagger, and user review. The English sentiment was translated to Indonesian automatically using set of seed word and expand them using sentiment patterns that found in the users review. The data was gathered from various of source such as Twitter, TripAdvisor and Openrice, the result shows that the proposed method can generate additional lexicon with sentiment accuracy of 77.7%.

These lexicon-based sentiment analysis is not limited by latin word just like in research by Abdulla et al [13] the research main purpose is comparing the two sentiment analysis approaches between the lexicon method and the corpus method on twitter using Arabic. For the corpus approach, SVM, Naïve Bayes, K-Nearest Neighbor and Decision Tree are combined with two Stemming processes: Root Stemmer and Light Stemmer. Next to lexicon approach used Clustering method where the test will be divided into 3 stages where each stage will be increased the amount of data. The results obtained for corpus approach get the highest result on SVM method with Light Stemmer with 87.2% accuracy. Then for the lexicon approach there is an increase in accuracy every increase in the amount of data where in the final stage obtained results 59.6% and allows accuracy will continue to increase if the amount of data has a lot. Lizhen [14] had a different approach by using Chinese language as the main object for the research , the research proposed new method which combines context-sensitive sentiment lexicon and the existing sentiment word list such as "Hownet" to enhance the domain specific opinion data analyzing performance. The raw dataset was crawled from 360buy.com and for the comparison the research was using Hownet based sentiment word list only and Clustering based sentiment only as the comparison. Result shows that the proposed method can give a better result under each measure such as Precision 75.02%, Recall 70.018%, and F-Measure 72.52%.

Muljono [15] conducting research of Indonesian text emotion detection and evaluates the performance with four different methods Naive Bayes, J48, KNN and SVM. The data containing 1000 sentences which consist six different emotion classes : anger, disgust, fear, joy, sadness, and surprise. The preprocessing method consist of tokenization case, normalization, stopword removal, stemming and TFIDF, with evaluation using 10-fold cross validation and split validation for the experiment. The result showing that the highest accuracy came from SVM with 85.5% with 10-fold cross validation and 86% from the split validation.

III. Research Method

In this chapter we will explain the methodology and the process steps that will be done in the research, see in figure 1 the initial stage is to obtain the dataset which will be used, then the process continue in preprocessing on the data that has been obtained which in this process is also used lexicon techniques to produce better data result, then we conduct the term weighting on the results of data that has been in preprocessing before, then after the research is done we will conduct comparison of classification test and reviewed the evaluation to get the best accuracy results from the comparison of classification methods that exist in this study.

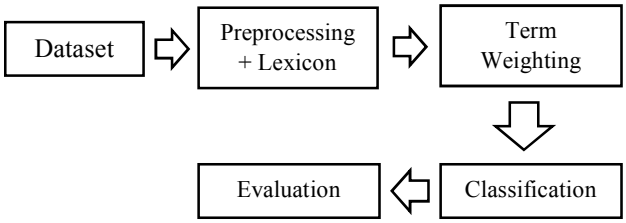


Fig. 1. Flow of Process Methods

A. Dataset

In this study will be used dataset review (response) from consumer amazon.com where this dataset obtained from UCI Machine Learning [16]with the amount of 1000 data, the data contains about the positive responses with the number 522 and negative responses amounted to 478. Later this dataset will be processed first using preprocessing so that data can be broken down and further processed in classification. This dataset in its attribute is nominal while for label column is numerical.

B. Preprocessing

In the Figure 2 we can see for this stage it will be used some features on preprocessing such as :

- Tokenize, this a feature used for extracting the sentences into a word from the dataset.
- Filter Stopword, this feature aim is for eliminate word that have no mention or purpose like “an”, “the”, etc.
- Filter Token (Length), Serves to eliminate words based on the number of letters, for example in this study the number of letters obtained only words amounted from 3 to 25 letters.
- Lexicon,at this stage there are three techniques in it that is stem (WordNet) which aims to change the affixed word become basic, then the technique of open WordNet Dictionary to help Stem (WordNet) techniques in choosing and changing the word, and last is Extract Sentiment which aims to extract the sentiment word to become weight.

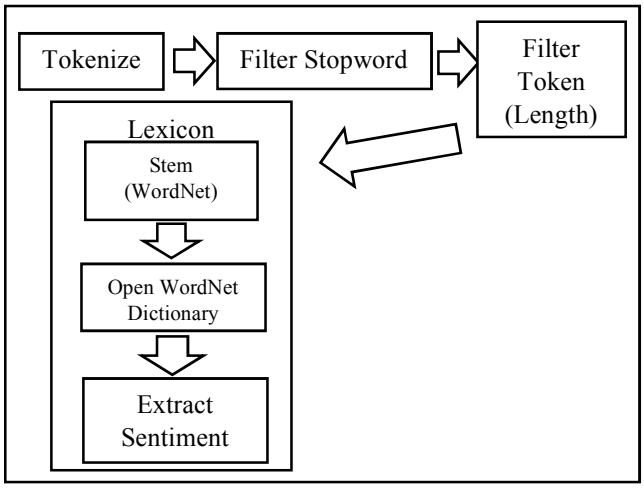


Fig. 2. Flow of Preprocessing

C. Term Weighting

In the process of Term Weighting will be used TF-IDF algorithm. The TF-IDF algorithm consists essentially of two algorithms, the Term Frequency (TF) algorithm, which is the frequency of occurrence of terms in related documents. The formula on Term Frequency is:

$$TF = \begin{cases} 1 + log10\,(ft,d), & ft,d > 0 \\ 0, & ft,d = 0 \end{cases} \quad (1)$$

Where the value of (ft, d) is the term frequency (t) of the document (d) [17].Then Inverse Document Frequency (IDF) is a calculation of the way term to be widely distributed in the collection of related documents. IDF is formulated as follows:

$$IDF_i = \log(D/df_j) \quad (2)$$

The variable D above is the sum of all documents in the collection, for (dfj) which is the number of documents containing term (tj) [17].If the above formula is combined it will be TF-IDF formulated as follows:

$$W_{ij} = tf_{ij}x\ \log(D/df_j) + 1 \quad (3)$$

Basically the value of (tfij), if D = dfj then the value will remain zero due to the result of log 1 for IDF calculation. So the solution of the problem taken is to add 1 value on the IDF side [17].

D. Classification

In order to find the best accuracy then in this research will use three classification methods as a comparison of the method of Decision Tree, K-Nearest Neighbor, and SVM. Here is an explanation of each classification method that will be used in this study.

1) Decision Tree

Decision Tree is a practical method of classifying data by trying to find a noise-resistant classification model. One of the most popular methods of Decision Tree is the ID3 method in which this method builds the decision tree model with top down by performing attribute evaluation using a statistic-based measure that usually takes the form of information gain, which will then be used to measure attribute attribute magnitude in the data classification [18].One method of decision tree is the C4.5 method using two calculations of Entrophy and Gain. Here is the formula of the two calculations are.

$$Entropy\ (S) = \sum_{j=1}^{k} Pj\ log2\ Pj \quad (4)$$

In the entropy calculation above S is the set of cases whereas k is the number of partitions on S and Pj is the probability obtained from the number of labels divided by the total problem / case [19].Then after finding entropy the next process will be calculated to find the gain formulated as follows.

$$Gain\ (S, A) = E\ (s) \sum_{i=1}^{n} -\frac{|Si|}{|s|} * E\ (si) \qquad (5)$$

In the above gain formula A is an attribute, then E is the entropy already obtained, n is the number of partition attributes of A, Si is the number of cases of the partition and s is the number of cases in the set [19].

2) K-Nearest Neighbor (K-NN)

It is a classification method that works in search of "k" data objects or patterns that have a position closest to the input pattern, then selects the class with the largest number of patterns among the k [18].

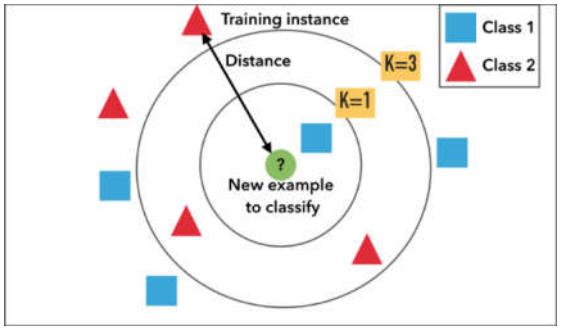


Fig. 3. Example Model K-Nearest Neighbor

In calculating two variables KNN it can use Pythagorean theorem where the distance between points x1 and x2 is calculated as follows:

$$d(x_1, x_2) = \sqrt{a^2 + b^2} = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2} \qquad (6)$$

In the above formula xij denotes the coordinates of point xi on the j dimension. That distance in mathematics is known as Euclidean Distance. By calculating the distance of the instance it can be done the instance distribution as the point and variable value j as the j-dimensional coordinate of the equation. As for instances that have more than two variables, Euclidean Distance can be generalized to many dimensions that are formulated as follows:

$$d(x_1, x_2) = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2 + \cdots + (x_{1p} - x_{2p})^2} = \sqrt{\sum_{j=1}^{p} (x_{1j} - x_{2j})^2} \qquad (7)$$

3) Support Vector Machine (SVM)

SVM is a learning methods that analyze data and recognize patterns which are used for classification and regression analysis [20]. The original SVM algorithm was created by Vladimir Vapnik and the current standard derivative (soft margin) proposed by Corinna Cortes and Vladimir Vapnik [21].

Basically SVM is a method to make predictions, it can be in the case of classification or regression. This method has similarities with Artificial Neural Network in the form of function and condition of problems to be solved so that SVM is also included in the class of supervised learning [22].This method separates data using hyperplane (one dimension less than the feature space, i.e a line with two dimensions) with the reference to it's closest data points called support vector [21]. What is the SVM tries is to maximize the support vector to get the maximum margin possible between the support vector and the hyperplane.

In the figure 4 a default SVM will fail if there is no hyperlane that separates the two classes which are the soft-margin SVM will choose a hyperplane that allows for the cleanest separation possible [20]. Therefore the soft-margin SVM has less chance of failing input that hard to separate but it may produce suboptiomal result where otherwise the default SVM would have failed.

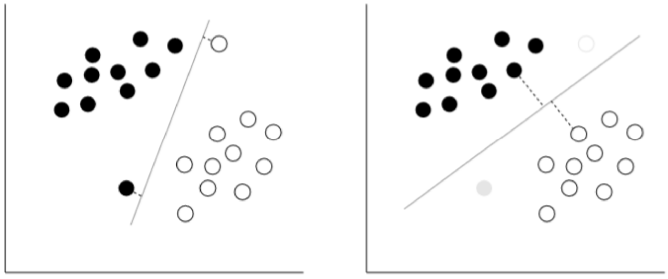


Fig. 4. Example Model of Default SVM (left) and Soft-Margin SVM (right)

This method is relatively new when compared to other methods, but SVM performs well in applications such as bioinformatics, handwriting recognition, text classification and so on [23]. According to Joachims [24] SVM has several advantages which are :

- Easier to handle without requiring special handling since most text are linearly separable.
- Can handles high-dimensional feature vector well.
- Well suited for sparse data since the feature vector are sparse because only a fraction of all words are used for each text.

IV. Result and Discussion

In this chapter we will describe the research done to find the best accuracy for sentiment analysis on amazon.com review. here are the steps in this research process.

A. Dataset

The data used in this study is data from product reviews on amazon.com obtained from UCI Machine Learning amounted to 1000 records. The data consisted of reviews and sentiments, reviews representing the content of customer responses and sentiments were the result of the assessment of Kotzias [16].Whether the response is positive or negative.

B. Figures and Tables

The aim of this research is to find the best accuracy of Decision Tree (DT), K-Nearest Neighbor (K-NN) classification method, and Support Vector Machine (SVM) by comparing the addition of lexicon feature during preprocessing process. The application used in this research is RapidMiner where this application can calculate all the method of classification in research. Evaluation used in this research is Split Validation which can do data sharing training and data testing by arranging ratio in RapidMiner application. The ratio used in the research is 0.7 which can be interpreted 70% data as training and 30% data will be testing on this test. Table I explain a comparison of results from classification calculations without lexicon:

TABLE I. ACCURACY METHODS WITHOUT LEXICON

Methods	Accuracy
Decision Tree	74.67 %
SVM	91.00 %
K-Nearest Neighbor	90.33 %

The highest result without Lexicon is 91% accuracy obtained from SVM then followed by K-NN with different this margin which is 90.33% and at last position there is Decision Tree which have accuracy that lower than any of the previous method that is 74.67%. From the acquisition of accurate results, SVM and K-NN actually already have good accuracy, but researchers will try to add lexicon feature on preprocessing where expected by adding lexicon feature can improve the accuracy of the classification method tested in the research. Table II explain a comparison of the classification methods tested with the addition of lexicon features to preprocessing.

TABLE II. ACCURACY METHODS WITH LEXICON

Methods	Accuracy
Decision Tree	82 %
SVM	91.33 %
K-Nearest Neighbor	92.67 %

With the addition of Lexicon feature on preprocessing can be seen in table II. There is an increase in each method of classification. The previous K-NN in second position is now in the first position with an accuracy of 92.67% with an increase of 2.34%. Then followed by SVM which previously ranked first position down to second position with 91.33 accuracy where only 0.33% increase so that can be overtaken by K-NN. In the last position is still occupied by the Decision Tree which despite a high increase but still can not follow the previous two classification methods. The accuracy obtained by the Decision Tree is 82% with an increase in accuracy of 7.33%.

V. Conclusion and Future Work

In this study we used the Decision Tree (DT) classification, K-Nearest Neighbor (K-NN), and Support Vector Machine (SVM) to find the best accuracy of the dataset we tested the data using Rapidminer tool. The ratio used in the research is 0.7 which can be interpreted 70% data as training and 30% data as testing on this test. The setting determines the ratio to the data, then we classified it, with SVM obtaining the highest result with 91% accuracy then K-NN with a slightly different result gain of 90.33% and at the last position there is a Decision Tree that has the lowest accuracy compared with the other two by 74.67%. We add lexicon feature to preprocessing, when result has variety of different result, best accuracy result can be by K-NN method with accuracy value 92.67% then SVM with 91.33% accuracy and Decision Tree with 82% accuracy. the addition of features to preprocessing is very influential on the classification, because the extraction of data sentiment so that found different results of each method of classification.

In the future research we would like to make some improvements into more realistic classification to the positive and negative impression that are expressed in the same review. Most lexicon-based have propose to make a neutral sentiment polarities when there is an opposite value occur in the same sentence and this does not seem to give good result. For example a word like “this movie is great but it has long screentime” is not considered as positive or negative that is why by most lexicon based method this is considered as neutral, but for many reason however this may far from becoming a neutral opinion. The one look text region in the polarity value for word, sentence, and text may be the source of this problem. a more sophisticated NLP techniques and more consideration on smaller text chunks will gonna be needed for handling them. Machine learning technique will also be considerable needed to classifying opionion in the amazon movie review and the result will be showed here.

References

[1] D. Lankelevich, "e-Marketer," 2016. [Online]. Available: www.emarketer.com.

[2] B. Jopson, "Amazon urges California referendum on online tax," 12 July 2011. [Online]. Available: FT.com.

[3] A. SAepuloh, "Catatan AepBae," 22 Maret 2013. [Online]. Available: www.aep-saepuloh.com.

[4] Muljono, D. P. Artanti, A. Syukur, A. Prihandono and D. R. . I. M. Setiadi, "Analisa Sentimen Untuk Penilaian Pelayanan Situs Belanja Online Menggunakan Algoritma Naïve Bayes," in *Konferensi Nasional Sistem Informasi STMIK*, Pangkalpinang, 2018.

[5] B. Pang, L. Lee and S. Vithyananthan, "Sentiment Classfication Using MAchine Learning Technique," in *Procedings of the ACL-02 conference on Empirical methods in natural language processing*, Stroudburg, 2002.

[6] R. Feldman and J. Sanger, The text mining handbook : advanced approaches in analyzing unstructure data, New York: Cambridge University Press, 2007.

[7] J. Shethna, "EDUCBA," 9 December 2016. [Online]. Available: www.educba.com/text-mining/.

[8] R. S.S.K., R. Singhal and A. Kumar, "An Improved Sentiment Classification using Lexicon into SVM," *International Journal of Computer Applications,* vol. 95, pp. 37-42, 2014.

[9] O. Kolchyna, . T. T. P. Souza, P. C. Treleaven and T. Aste, "Twitter Sentiment Analysis: Lexicon Method, Machine Learning Method and Their Combination," *CEUR Workshop Proceedings,* pp. 1-32, 2015.

[10] L. V. Avanco and M. G. V. Nunes, "Lexicon-based Sentiment Analysis for Reviews of Products in Brazilian Portuguese," in *Brazilian Conference on Intelligent Systems*, Sao Paolo, 2014.

[11] A. F. Wicaksono, C. Vania, B. D. T and M. Adriani, "Automatically Building a Corpus for Sentiment Analysis on Indonesian Tweets," in *Pacific Asia Conference on Language, Information and Computation* , 2014.

[12] C. Vania, M. Ibrahim and M. Adriani, "Sentiment Lexicon Generation for an Under - Resourced Language," *IJCLA,* vol. 5, pp. 59-72, 2014.

[13] N. A. Abdulla, N. A. Ahmed, M. A. Shehab and M. Al-Ayyoub, "Arabic Sentiment Analysis: Lexicon-based and Corpus-based," *IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies,* pp. 1-6, 2013.

[14] L. Liu, M. Lei and H. Wang, "Combining Domain-Specific Sentiment Lexicon with Hownet for Chinese Sentiment Analysis," *Journal of Computers,* vol. 8, pp. 878-883, 2013.

[15] Muljono, A. S. Winarsih and C. Supriyanto, "Evaluation of Classification Methods for Indonesian Text Emotion Detection," in *ISEMANTIC*, 2016.

[16] D. Kotzias, "UCI Machine Learning Repository," 30 May 2015. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Sentiment+Labelled+Sentences.

[17] INFORMATIKALOGI, "informatikalogi.com," 11 July 2017. [Online]. Available: https://informatikalogi.com/term-weighting-tf-idf/.

[18] S. S.T, Data Mining untuk Klasifikasi dan Klasterisasi Data, Bandung: Penerbit Informatika, 2017.

[19] Marwana, "Algoritma C4.5 Untuk Simulasi Prediksi Kemenangan Dalam Pertandingan Sepakbola," *Jurnal Informatika Multimedia, STIMED Nusa Palapa.*

[20] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, no. 3, pp. 273-297, 1995.

[21] F. Sommar and M. Wielondek, Combining Lexicon- and Learning-based Approaches for Improved Performance and Convenience in Sentiment Classification, Stockholm, 2015.

[22] S. Santoso, Buku Latihan SPSS Statistik Parametik, Jakarta: PT. Eleks Media Komputindo, 2007.

[23] B. Ramana, S. P. Babu and N. B. Venkateswarlu, "A Critical Study Of Selected Classification Algorithms For Liver Disease Diagnosis," *International Journal Of Database Management Systems ,* vol. 3, no. 2, pp. 101-114, 2011.

[24] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," *Springer,* 1998.

CBIR for Herbs Root using Color Histogram and GLCM based on K-Nearest Neighbor

Candra Irawan, Winda Listyaningsih, De Rosal Ignatius Moses Setiadi, Christy Atika Sari, Eko Hari Rachmawanto
Department of Informatics Engineering, Computer Science Faculty
Dian Nuswantoro University
Semarang, Indonesia
Email:candra.irawan@dsn.dinus.ac.id, windalistyaningsih25@gmail.com, moses@dsn.dinus.ac.id, atika.sari@dsn.dinus.ac.id, eko.hari@dsn.dinus.ac.id

***Abstract*— Ginger, *kencur*, *kunyit*, *temulawak*, and *temu hitam* are herbs root that is widely used as a spice and herbal medicine in Indonesia. Visually the shape and color of herbs root are quite similar so that the common people difficult distinguish. The problem can be solved with the help of Content-Based Image Retrieval (CBIR). There are three main modules in CBIR namely pre-processing, feature extraction and identification. The study proposed the Color Histogram and Gray Level Co-occurrence Matrix (GLCM) as feature extraction to identify these types of herbal remedies. While the identification stage used K-Nearest Neighbor (KNN) and the calculation of the distance closest to the Euclidean distance. The amount of data used as training data is 250 images, while data testing used 25 images. The results of this study the system can recognize the image of herbs root in accordance with the type with the highest accuracy obtained.**

Keywords— Herbs Root, KNN, CBIR, GLCM, Color Histogram

I. INTRODUCTION

Herbs and spices are one of Indonesia's natural wealth, which is widely used as flavoring herbs and herbal remedies. All parts of medicinal plants can be used among others is the tuber. Examples of medicinal plants taken from tuber include ginger, *kencur*, *kunyit*, *temulawak*, and *temu hitam*. The five types of medicinal plants belonging to the Plantae class with the *Zingiberaceae* family [1]. For people who are not familiar, visually the shape and color of the root of this medicinal plant is almost the same and sometimes even need to smell the aroma to know the difference. Another way to work around this is to ask the botanists to identify the types of medicinal plants, which are inefficient. Therefore, we need a tool to automatically identify medicinal plants.

Content-Based Image Retrieval (CBIR) is one technique that can be used to help to identify it. CBIR is a data search technique with image input [2]. Several previous studies as having been done in Sutojo et al [3] using Color Histogram and Gray Level Co-occurrence Matrix (GLCM) on cow image with CBIR search technique using proximity distance count between Euclidean distance and evaluation images using precision, recall and accuracy values from Confusion Matrix table. Proven excellent with 95% accuracy, 100% precision, and recall. Another study conducted by Awaj et al [4] also used GLCM to classify the image of herbal leaves. GLCM feature extraction results were measured by the closest distance to Canberra method with the highest accuracy of 92%. Indriani et al [5] also use feature color extraction and feature extraction of GLCM textures to classify tomatoes into 5 classes ie raw, well-done, ripe, very ripe and rotten. The combined results of both feature extractions are classified by KNN and the distance measured nearest to the Euclidean distance yields perfect accuracy of 100% with k = 3.

This study proposes a combination of color and texture feature extraction to identify five types of image of medicinal plants, namely ginger, *kencur*, *kunyit*, *temulawak* and *temu hitam*. The Color Histogram is proposed as a color feature extraction and GLCM is proposed as a texture feature extraction. A color histogram is a graph that describes the difference in pixel image intensity value in certain parts [6]. GLCM is one method of texture extraction based on the degree of the gray image [7]. The identification stage is proposed using KNN and Euclidean distance methods to determine the closest similarity of the image.

II. COLOR HISTOGRAM

Color Histogram is a representation of the distribution of colors scattered in an image obtained by counting the number of pixels of each part of the color range [3]. There are two features used from the color histogram that is by calculating the average (μ) color and standard deviation (σ) the color intensity of each RGB color space. The average (μ) color is used to determine the level of color depth in an image that can be calculated by the formula (1).

$$\mu = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N} X_{i,j} \quad (1)$$

$$\sigma = \sqrt{\frac{1}{(MN)-1}\sum_{i=1}^{M}\sum_{j=1}^{N}(X_{(i,j)} - \mu)^2} \quad (2)$$

Where μ = mean, σ = standard deviation, MN = size of row and column image, X_{ij} = image pixel value (X) on row i and column j. While the standard deviation (σ) color intensity is used to know the distribution of color in an image.

III. GRAY LEVEL CO-OCCURRENCE MATRIX (GLCM)

GLCM is a technique used to obtain a second order statistic value by calculating the probability of the proximity relationship between two pixels at a certain distance (d) and angle (θ) [7]. The process performed on GLCM calculations using a co-occurrence matrix. In determining the pairs between neighboring pixels, GLCM uses four commonly used angles: 0 °, 45 °, 90 ° and 135 ° [8]. Commonly, the distance used is one pixel. In GLCM calculations use five features, namely energy, contrast, homogeneity, entropy, and correlation. Energy is the sum of the uniformity of color in an image [8], where this feature can be calculated by the formula (3). The contrast reflects the clarity of an image [9], where this feature can be calculated by the formula (4). The contrast will be 0 if the pixel has a similar value. Homogeneity is a measure of the similarity of the variation of the co-occurrence matrix in the image [3], where this feature can be calculated

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

by the formula (5). Entropy is a measure of the gray level in an irregular image with the variation of each element of the co-occurrence matrix. Images with large gray degrees have large variation values [10], where this feature can be calculated by the formula (6). Correlation is a measure of linear dependence between the gray level values in the image [11], where this feature can be calculated by the formula (7).

$$energy = \sum_{i=1}^{M}\sum_{j=1}^{N} X_{(i,j)^2} \quad (3)$$

$$contrast = \sum_{i=1}^{M}\sum_{j=1}^{N} (i,j)^2 X_{i,j} \quad (4)$$

$$homogeinity = \sum_{i=1}^{M}\sum_{j=1}^{N} \frac{X_{i,j}}{1+|i-j|} \quad (5)$$

$$entropy = -\sum_{i=1}^{M}\sum_{j=1}^{N} X_{ij} log^2 X_{ij} \quad (6)$$

$$correlation = \sum_{i=1}^{M}\sum_{j=1}^{N} \frac{(i-\mu_i)(j-\mu_j)}{\sigma_i \sigma_j} \quad (7)$$

Where MN = row size and image column, X_{ij} = image pixel value (X) on row i and column j, μ = mean (mean), σ = standard deviation, and

$$\mu_i = \sum_i \sum_j i X_{(i,j)} \quad (8)$$

$$\mu_j = \sum_i \sum_j j X_{(i,j)} \quad (9)$$

$$\sigma i = \sqrt{\sum_{i}^{M}\sum_{j}^{N} (i-\mu_i)^2 X_{(i,j)}} \quad (10)$$

$$\sigma j = \sqrt{\sum_{i}^{L}\sum_{j}^{L} (j-\mu_j)^2 X_{(i,j)}} \quad (11)$$

IV. PROPOSED METHOD

The following is an explanation of the stages of the proposed research to search the image of tuber bulbs with CBIR. Based on figure 1 can be described in more detail as follows:

A. Dataset

Dataset consists of training image and image testing.

B. Preprocessing

At this stage, all datasets are done three processes that change the background with white color. Next crop with the size adjusted to the largest object size taken. Finally, resize images with a size of 400 x 600 pixels, to speed up the process of reading images.

C. Features Extraction

The next step is to extract the color features based on the RGB color space. The extraction of color features taken is the mean and deviation standard. Next, the image is converted to grayscale to extract its texture feature using GLCM. Extraction of texture features taken is of five kinds, namely energy, contrast, homogeneity, entropy, and correlation.

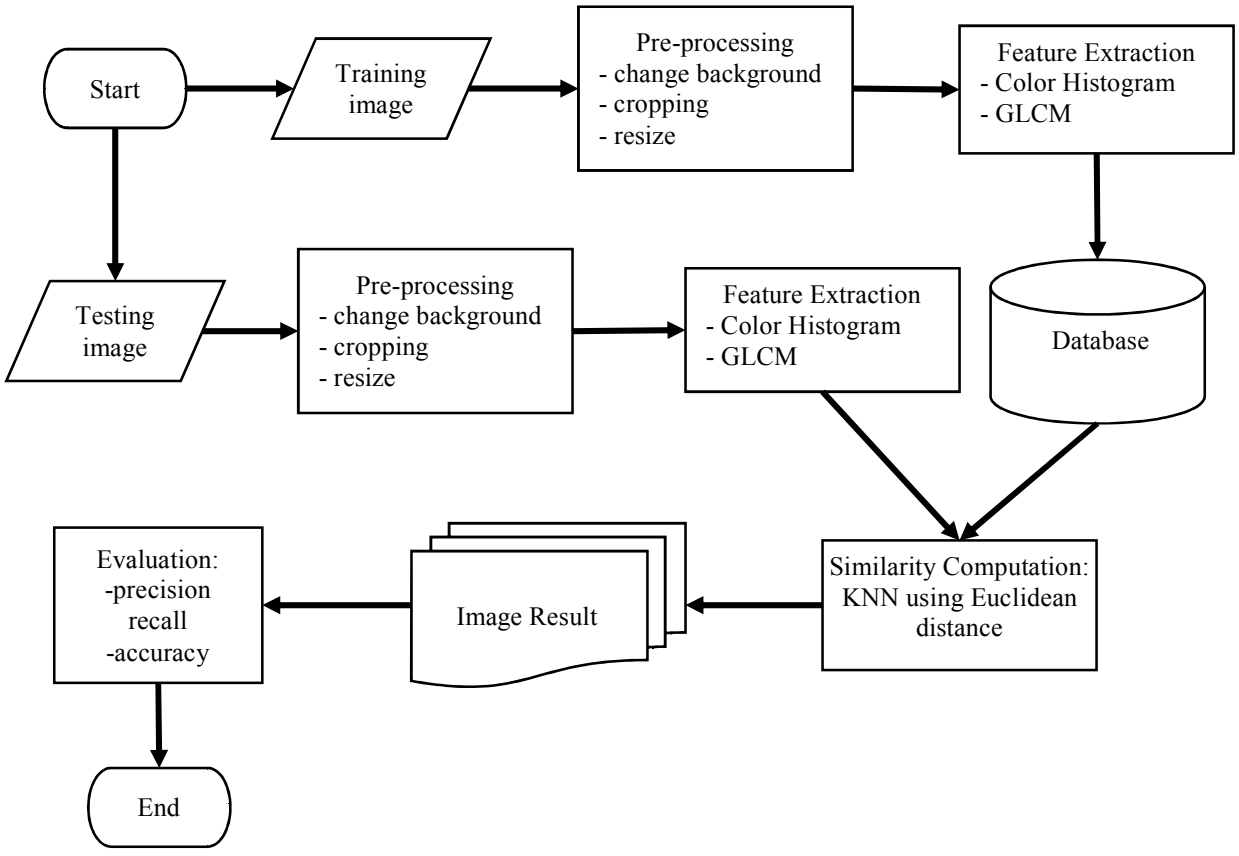


Fig. 1. The flowchart of the proposed method

D. Database

The database is used as a place to store the value of features that have been labeled on the entire training image.

E. Similarity Computation

The level of similarity between these images was calculated using the Euclidean distance measurement method. The value of the train image that has the smallest distance with the test image means that the image is most similar to the test image. The formula used to calculate the Euclidean distance using formula (12).

$$Euclidean = \sqrt{\sum_{k=1}^{n}(X_{ik} - X_{ij})^2} \quad (12)$$

F. Image Result

In the search results will be displayed in the first ten training images that have the closest distance to the test image

G. Evaluation

The performance of the proposed method is measured using a confusion matrix by calculating the value of accuracy, precision, and recall.

V. Experiment and Results

A. Dataset

Image dataset used five types of images of ginger, *kencur*, *kunyit*, *temulawak*, and *temu hitam*. Before the image is taken from the five types of medicinal plants are washed clean first. Next image is taken using rear camera Lenovo Vibe K5 with 13-megapixel camera resolution. The number of images of each type taken 55 images, the total image obtained are 275 images, which consists of 25 test images and 250 train images. Furthermore the image is named according to the label, to shorten the naming of 250 pieces of training image used as a training image with details 1_1.jpg to 1_50.jpg are ginger images, 2_1.jpg to 2_50.jpg are *kencur* images, 3_1.jpg to 3_50.jpg are *kunyit* images, 4_1.jpg until 4_50.jpg are *temulawak* images and 5_1.jpg to 5_50.jpg are *temu hitam* images. Similarly in the test image is also used naming with the format.

B. Preprocessing

In the preprocessing process is done three stages change the background to white, cropping, and resize.

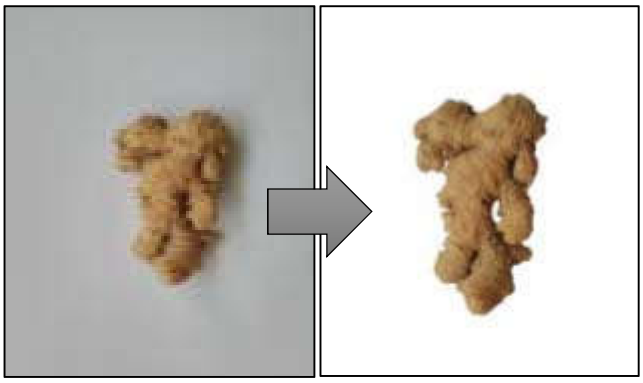


Fig. 2. The result of the background change process in the sample of the ginger image

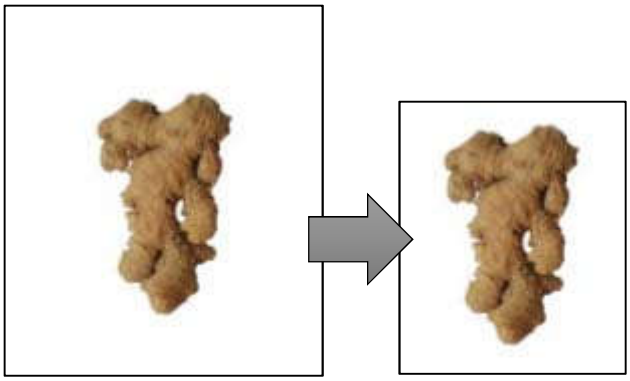


Fig. 3. Results of the cropping process on the sample of ginger image

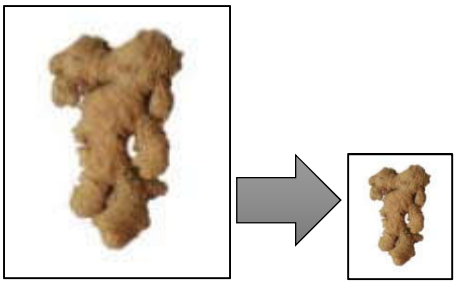


Fig. 4. Results of the resize process on the sample of ginger image



(a) (b) (c) (d) (e)

Fig. 5. Example of dataset image after preprocessing {(a) ginger, (b) *kencur*, (c) *kunyit*, (d) *temulawak*, (e) *temu hitam*}

C. Color Histogram

At this stage, the RGB color space is separated and then taken the mean and standard deviation values of each color space. Table 1 shows the extraction of color features based on the color histogram on the training image, while Table 2 shows the extraction feature value in the test image.

TABLE I. Mean and Standart Deviation Value of Data Training

Images	Mean			Standard deviation		
	Red	*Green*	*Blue*	*Red*	*Green*	*Blue*
1_1	241.73	237.04	231.76	37.26	49.03	62.04
1_2	242.27	237.77	232.92	37.35	48.82	61.24
…	…	…	…	…	…	…
2_1	245.00	241.43	239.13	37.26	48.62	56.04
2_2	238.80	231.52	226.80	43.70	60.69	71.61
…	…	…	…	…	…	…
3_1	246.75	243.74	242.33	30.91	41.41	46.30
3_2	245,88	243,09	241,47	33,52	43,142	48,77
…	…	…	…	…	…	…
4_1	245.89	243.10	241.47	33.52	43.14	48.77
4_2	217,13	203,19	192,10	55,608	74,56	89,47
…	…	…	…	…	…	…
5_49	234.26	228.00	221.95	47.08	60.22	73.00
5_50	231.97	225.26	217.99	48.26	61.48	75.76

TABLE II. MEAN AND STANDART DEVIATION VALUE OF DATA TESTING

Citra	Mean			Standard deviation		
	Red	*Green*	*Blue*	*Red*	*Green*	*Blue*
1_1	247.07	243.83	241.14	32.97	43.98	53.24
…	…	…	…	…	…	…
1_5	240.95	235.85	231.26	43.58	56.74	68.82
2_1	246.91	245.04	243.01	31.73	38.82	46.46
…	…	…	…	…	…	…
2_5	248.46	245.99	243.93	27.14	36.90	45.06
3_1	242.34	237.70	234.93	43.35	57.21	65.76
…	…	…	…	…	…	…
3_5	245.07	241.89	239.57	39.00	50.04	58.31
4_1	221.10	207.74	194.00	61.42	77.35	95.88
…	…	…	…	…	…	…
4_5	233.34	223.09	212.18	50.21	67.39	87.38
5_1	249.39	247.48	246.20	29.53	38.58	44.71
…	…	…	…	…	…	…
5_5	243.70	239.84	237.28	41.17	53.71	62.12

D. Gray Level Co-Occurrence Matrix (GLCM)

To perform feature extraction with GLCM the RGB image is converted to a grayscale image. Below is the result of the experiment with the calculation of the 5 features of energy, contrast, homogeneity, entropy, and correlation. The five features are taken from four angles, namely 0 °, 45 °, 90 ° and 135 °. Table III to Table VI shows the GLCM feature extraction results in the training data, whereas in tables VII to X shows the GLCM feature extract results in the data testing.

TABLE III. GLCM FEATURE VALUE OF DATA TRAINING AT 0 °

Images	Energy	Contrast	Homo	Entropy	Corr
1_1	0.75521	60.22107	0.88371	1.52203	0.00046
1_2	0.76862	61.06006	0.89012	1.43711	0.00046
…	…	…	…	…	…
2_1	0.84774	45.71135	0.92871	0.97016	0.00048
2_2	0.73526	52.25095	0.87734	1.64345	0.00031
…	…	…	…	…	…
3_1	0.85679	27.42450	0.93663	0.87024	0.00067
3_2	0.85331	30.95631	0.93432	0.89320	0.00060
...	…	…	…	…	…
4_1	0.67514	52.41814	0.84953	1.92110	0.00030
4_2	0.63690	53.59507	0.82355	2.18034	0.00029
…	…	…	…	…	…
5_49	0.86365	23.22393	0.93930	0.83326	0.00047
5_50	0.86693	25.93713	0.93957	0.82814	0.00050

TABLE IV. GLCM FEATURE VALUE OF DATA TRAINING AT 45 °

Images	Energy	Contrast	Homo	Entropy	Corr
1_1	0.75432	89.13882	0.88119	1.54233	0.00045
1_2	0.76722	110.50108	0.88713	1.46288	0.00046
…	…	…	…	…	…
2_1	0.84630	90.58254	0.92640	0.98912	0.00047
2_2	0.73422	100.65981	0.87159	1.68377	0.00031
…	…	…	…	…	…
3_1	0.85590	54.02105	0.93344	0.88885	0.00066
3_2	0.85262	51.28918	0.93161	0.91146	0.00060
...	…	…	…	…	…
4_1	0.67420	100.27050	0.84268	1.96593	0.00030
4_2	0.63543	108.35566	0.81685	2.22934	0.00029
…	…	…	…	…	…
5_49	0.86207	66.68002	0.93594	0.85653	0.00046
5_50	0.86519	75.47730	0.93619	0.85244	0.00050

TABLE V. GLCM FEATURE VALUE OF DATA TRAINING AT 90 °

Images	Energy	Contrast	Homo	Entropy	Corr
1_1	0.75507	54.43373	0.88468	1.51456	0.00046
1_2	0.76859	65.28510	0.89029	1.43597	0.00046
…	…	…	…	…	…
2_1	0.84748	51.46022	0.92851	0.97340	0.00048
2_2	0.73428	87.66116	0.87282	1.67470	0.00031
…	…	…	…	…	…
3_1	0.85669	31.89376	0.93500	0.87779	0.00067
3_2	0.85345	31.79698	0.93336	0.90037	0.00060
...	…	…	…	…	…
4_1	0.67481	67.75867	0.84579	1.94479	0.00030
4_2	0.63621	80.14912	0.81979	2.20698	0.00029
…	…	…	…	…	…
5_49	0.86213	59.94009	0.93698	0.84947	0.00046
5_50	0.86551	58.47715	0.93735	0.84375	0.00050

TABLE VI. GLCM FEATURE VALUE OF DATA TRAINING AT 135 °

Images	Energy	Contrast	Homo	Entropy	Corr
1_1	0.75398	93.78616	0.88105	1.54401	0.00045
1_2	0.76777	86.51946	0.88816	1.45339	0.00046
…	…	…	…	…	…
2_1	0.84716	67.28646	0.92711	0.98264	0.00047
2_2	0.73302	129.62411	0.87054	1.69355	0.00031
…	…	…	…	…	…
3_1	0.85610	46.23031	0.93357	0.88783	0.00066
3_2	0.85271	53.32624	0.93168	0.91044	0.00060
...	…	…	…	…	…
4_1	0.67404	98.72218	0.84207	1.97173	0.00030
4_2	0.63560	105.94639	0.81714	2.23051	0.00029
…	…	…	…	…	…
5_49	0.86174	77.36430	0.93592	0.85765	0.00046
5_50	0.86527	69.17687	0.93664	0.85036	0.00050

TABLE VII. GLCM FEATURE VALUE OF DATA TESTING AT 0 °

Images	Energy	Contrast	Homo	Entropy	Corr
1_1	0.86659	58.15517	0.93647	0.86526	0.00058
…	…	…	…	…	…
1_5	0.75792	63.46066	0.88473	1.50404	0.00044
2_1	0.87275	15.04044	0.94988	0.73210	0.00071
…	…	…	…	…	…
2_5	0.94452	6.62635	0.97774	0.33920	0.00157
3_1	0.82942	48.77173	0.92439	1.04482	0.00034
…	…	…	…	…	…
3_5	0.77490	44.26965	0.89825	1.38646	0.00026
4_1	0.48813	53.27470	0.75807	3.21227	0.00018
…	…	…	…	…	…
4_5	0.46946	106.43466	0.72508	3.38294	0.00014
5_1	0.92234	12.34193	0.96603	0.48881	0.00076
…	…	…	…	…	…
5_5	0.70742	53.85135	0.86144	1.82007	0.00025

TABLE VIII. GLCM FEATURE VALUE OF DATA TESTING AT 45 °

Images	Energy	Contrast	Homo	Entropy	Corr
1_1	0.86604	83.45038	0.93493	0.87543	0.00058
…	…	…	…	…	…
1_5	0.75647	114.61223	0.88170	1.53113	0.00044
2_1	0.87140	45.15789	0.94503	0.75769	0.00071
…	…	…	…	…	…
2_5	0.94374	21.38598	0.97596	0.34995	0.00155
3_1	0.82889	73.81103	0.92161	1.06319	0.00034
…	…	…	…	…	…
3_5	0.77388	89.54395	0.89274	1.42093	0.00026
4_1	0.48677	116.00502	0.74206	3.30832	0.00018

…	…	…	…	…	…
4_5	0.46824	231.28857	0.71138	3.49734	0.00014
5_1	0.92105	44.42051	0.96341	0.50478	0.00075
…	…	…	…	…	…
5_5	0.70349	180.09133	0.85260	1.88833	0.00025

TABLE IX. GLCM FEATURE VALUE OF DATA TESTING AT 90 °

Images	Energy	Contrast	Homo	Entropy	Corr
1_1	0.86676	49.77815	0.93699	0.86241	0.00058
…	…	…	…	…	…
1_5	0.75783	67.99315	0.88492	1.50383	0.00044
2_1	0.87174	35.27574	0.94724	0.74839	0.00071
…	…	…	…	…	…
2_5	0.94383	18.07888	0.97670	0.34669	0.00155
3_1	0.82974	37.49093	0.92443	1.04527	0.00035
…	…	…	…	…	…
3_5	0.77417	71.98496	0.89463	1.40823	0.00026
4_1	0.48715	93.16829	0.74828	3.27130	0.00018
…	…	…	…	…	…
4_5	0.46875	181.19304	0.71455	3.46854	0.00014
5_1	0.92107	42.62715	0.96419	0.50147	0.00075
…	…	…	…	…	…
5_5	0.70416	141.46755	0.85527	1.86480	0.00025

TABLE X. GLCM FEATURE VALUE OF DATA TESTING AT 135°

Images	Energy	Contrast	Homo	Entropy	Corr
1_1	0.86577	86.40725	0.93544	0.87428	0.00058
…	…	…	…	…	…
1_5	0.80202	75.76378	0.90333	1.28178	0.00091
2_1	0.75707	89.77920	0.88248	1.52055	0.00044
…	…	…	…	…	…
2_5	0.92602	26.78474	0.96981	0.43446	0.00136
3_1	0.94364	22.01061	0.97601	0.35013	0.00155
…	…	…	…	…	…
3_5	0.86049	71.19260	0.93458	0.88375	0.00041
4_1	0.77382	99.97921	0.89216	1.42391	0.00026
…	…	…	…	…	…
4_5	0.57636	85.79684	0.80673	2.57215	0.00015
5_1	0.46817	226.85639	0.71214	3.49246	0.00014
…	…	…	…	…	…
5_5	0.70405	155.12132	0.85376	1.87698	0.00025

E. Similarity Computation

Similarity Computation is a calculation used to calculate the similarity between the test image and the training image uses the Euclidean distance in the formula (12). The calculation results are shown in Table XI.

TABLE XI. EUCLIDEAN DISTANCE OF DATA TESTING

Images	Energy
1_1	38.8462114
…	…
1_5	60.18053575
2_1	50.52645863
…	…
2_5	50.52645863
3_1	93.16436511
…	…
3_5	35.64791986
4_1	117.6415262
…	…
4_5	286.6831244
5_1	44.37784612
…	…
5_5	166.8620869

F. Image Result

The image displayed in the search results column is the first 10 images that have the smallest distance. The smallest distance can be said that it has the similarity level closest to the testing image. Figure 6 shows sample search results on the ginger image (1_1.jpg).



Fig. 6. Image search results 1_1.jpg

G. Evaluation

In this evaluation process calculate the three values, namely accuracy, precision, and recall based on the tested image using the confusion matrix table shown in Table XII.

TABLE XII. CONFUSION MATRIX TABEL

		Prediction Value				
		1	2	3	4	5
Real Value	1	5				
	2		4			1
	3		1	2		2
	4				5	
	5	2			1	2

Explanation:
1 = Ginger.
2 = *Kencur*.
3 = *Kunyit*.
4 = *Temulawak*.
5 = *Temu Hitam*.

$$Accuraccy = \frac{5+4+2+5+2}{25} = 72\%$$

$$Precision = \frac{5+4+2+5+2}{18} = 100\%$$

$$Recall = \frac{5+4+2+5+2}{18+0} = 100\%$$

VI. CONCLUSION

Ginger, *kencur*, *kunyit*, *temulawak*, and *temu hitam* is a herbs root that is widely used as a spice and herbal medicine in Indonesia. Visually the shape and color of herbs root are quite similar so that the common people difficult distinguish. This study proposes CBIR using color histogram color feature extraction and KNN-based GLCM texture feature extraction to identify herbs plant root. Based on experimental results it is evident that the proposed method has the highest accuracy of 72%. This result is quite good considering visually these five types of plants look similar. In the next research, accuracy needs to be optimized by adding the amount of training data and the preprocessing stage or by applying extraction features and can also use several combination methods.

REFERENCES

[1] S. Hidayat, Kitab Tumbuhan Obat, 1st Edition ed., F. A. Nurrochman, Ed., Cibubur: AgriFlo, 2015.

[2] Kannan, "Content Based Image Retrieval Using Nearest Neighbour and Hybrid KNN SVM Methods to Diagnose MR Images," *International Journal of Innovative Research in Science, Engineering, and Technology,* vol. III, no. 3, pp. 3021-3026, March 2014.

[3] T. Sutojo, P. S. Tirajani, D. R. I. M. Setiadi, C. A. Sari and E. H. Rachmawanto, "CBIR for Classification of Cow Types using GLCM and Color Features Extraction," in *International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Yogyakarta, 2017.

[4] M. F. Awaj, "Pattern Recognition Analysis of Herbal Leaves Based on Distance-Based Similarity Measures Using the Gray Level Co-Occurrence Matrix Feature Extraction," in *International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*, Semarang, 2017.

[5] O. R. Indriani, E. J. Kusuma, C. A. Sari, E. H. Rachmawanto and D. R. I. M. Setiadi, "Tomatoes Classification Using K-NN Based on GLCM and HSV Color Space," in *International Conference on Innovative and Creative Information Technology (ICITECH)*, Salatiga, 2017.

[6] R. K. Lingadalli, "Content Based Image Retrieval using Color, Shape, and Texture," *International Advanced Research Journal in Science Engineering and Technology,* vol. 2, no. 6, pp. 40-45, June 2015.

[7] K. Gaurav, A. Verma, H. Kumar and R. Hakur Anupam Mukherjee, "Content Based Image Retrieval using GLCM," *International Journal of Innovative Research in Computer and Communication Engineering,* vol. IV, no. 11, pp. 20142-20149, November 2016.

[8] R. A. Pramunendar, C. Supriyanto, D. H. Novianto, I. N. Yuwono, G. F. Shidik and P. N. Andono, "A classification method of coconut wood quality based on Gray Level Co-occurrence matrices," in *International Conference on Robotics, Biomimetics, Intelligent Computational Systems*, Jogjakarta, 2013.

[9] D. John, S. Tharani, and S. K, "Content Based Image Retreival using HSV-Color Histogram and GLCM," *International Journal of Advance Research in Computer Science and Management Studies (www.ijarcsms.com),* vol. II, no. 1, pp. 246-253, January 2014.

[10] A. M. Bhave and M. Wanjari, "IRetrieval : Image Retrieval Based on Color Feature and Texture Feature," *International Journal of Advanced Research in Computer Science,* vol. V, no. 6, pp. 165-167, Juli 2014.

[11] A. Saini and S. Singh, "Evaluation of CBIR System by Using Low Level Features of an Image," *International Journal of Science, Engineering and Technology Research (IJSETR),* vol. V, no. 6, pp. 1951-1954, June 2016.

[12] P. Mohanaiah, P. Sathyanarayana, and L. GuruKumar, "Image Texture Feature Extraction Using GLCM Approach," *International Journal of Scientific and Research Publications,* vol. 3, no. 5, pp. 1-5, 2013.

IMAGE ENHANCEMENT SEGMENTATION AND EDGE DETECTION IN MRI FOR MAMMOGRAM DISEASE

Moch.Syahrir[1], Akhmad Kuncoro[2], Aldhi Ari Kurniawan[3], M.Arief Soeleman[4*], Guruh Fajar Shidik[5*]

[1,3] *Department of Informatics FT UDINUS*

Imam Bonjol Street No. 207, Pindrikan Lor, Semarang Tengah, Semarang city, central java 50131

[1]Moch.Syahrir muhammadsyahriralfath@gmail.com

[2]Akhmad Kuncoro akh.kuncoro@gmail.com

[3]Aldhi Kurniawan aldhiarikurniawan@gmail.com

[4*]M. Arief Soeleman arief22208@gmail.com

[5*]Guruh Fajar Shidik guruh.shidik@dsn.dinus.ac.id

Abstract-Magnetic resonance imaging (MRI) is a medical imaging technique used in radiology to visualize internal structures in detail. MRI technology is excellent for identifying cancers, in identifying breast cancer, usually doctors or radiologists analyse images of magnetic resonance stored in Digital Imaging Communication in Medicine (DICOM) format. It takes expertise and experience enough for the diagnosis given precisely and accurately, so that handling can be done wisely. Image segmentation is intended to get the objects contained within the image and separate it from the background, while the edge detection is done to obtain the edges of the object. However, edge detection results are sometimes still not maximal therefore after edge detection is done to improve the image or commonly called image enhancement, to improve the image can be done in various ways tailored to the needs, in this study carried out by the High Pass Filter approach. Stages of the process consists of Gray scaling, Tresholding, Cropping, Edge Detection with Prewitt Operator, Image Enhancement then the image of the Enhancement result is done by Union Operation to fill the white pixel value which is in Edge Detection and then converted to binary value 0 and 1. This study used 4 test images consisting of 2 normal images and 2 abnormal images, 2 normal images in the mean as sampling. The test results with edge detection is not good enough so after edge detection is done enhancement to the image so as to produce the image with the appearance of the boundary line of each edge of the area is clearly visible and there is no disconnected edge

Keywords – MRI, Edge Detection, Operator Union, Image Enhancement, High Pass Filter

I.INTRODUCTION

The development of technology in the field of increasingly sophisticated health by using tools that modern and sophisticated prove that the health of the world progress. One application of technology in the field of health that is currently widely used is Magnetic Resonance Imaging (MRI) [1]. One type of cancer that can be identified by using this MRI technology is breast cancer. Breast cancer is still the leading cause of death of the world.

In identifying breast cancer that is stored in Digital Imaging Communication in Medicine format. Research to identify breast cancer using medical images of Magnetic Resonance results is still rare, because research by utilizing medical images of MRI results more aimed at identifying brain cancers in humans [1]. In addition, based on a doctor or radiological statement that from the results of an MRI examination cannot be diagnosed with certainty the presence of breast cancer, further clinical examination is necessary to ensure the patient [1].

Magnetic Resonance Imaging (MRI) is an advanced diagnostic tool for checking and detecting the body using large magnetic fields and radio frequency waves, without surgery, use of X-rays, or radioactive material, resulting in recording images of cross-section of the body or human organs using magnetic field strength between 0,064-, 5 tesla (1 tesla = 1000 Gauss) and resonance of vibration to nucleus of hydrogen atom

Image segmentation is a process aimed at obtaining objects contained within the image or dividing the image into several regions with each object or region having similar attributes [2]. In an image containing only one object, the object is distinguished from its background. The image segmentation technique is based on two basic properties of gray-level values: incompatibility and similarity between pixels [3].
Edge detection serves to obtain the edge of the object [4]. Edge detection takes advantage of drastic changes in the intensity value of the boundaries of two areas [5]. The definition of the edge here is "the set of connected pixels that lie on the boundaries of two areas" (Gonzales & Woods 2002)

[12]. Need to know the real edge contains very important information. The information obtained can be either the shape or size of the object. Generally edge detection uses two kinds of detector ie detecktor line (Hy) and column detector (Hx) [4]. Some examples of this type are operators Robert, Prewitt, Sobel, Canny, and Frei Chen.

Edge detection can be divided into two groups [4]. The first group is called first-order edge detection, which works by using first order derivative or differential [1]. Included in this group are the operators Roberts, Prewitt, and Sobel. The second group is called second-order edge detection, which uses second-order derivatives [6]. An example belonging to this group is Laplacian of Gaussian.

In previous studies, sometimes a study refers only to segmentation and edge detection to know the regularity pattern of an object in an image, but sometimes the image we get from edge detection is not always good, sometimes there is a lack of clear image, or not connected or otherwise, for this problem we can perform image repairs in various ways tailored to the needs of one of them with High Pass Filter approach aim to revive the edge of the image that has been produced by edge detection operations, and also to reduce the analysis of human work, image results from edge detection we can do again so that we can easily take a decision from the analysis. One that can be used is a union operation.

So simply this research is done in three stages of image edge detection will be repaired the edges, because the edge detection white pixels only at the edges only. Then proceed with morphological operations that focus only on union surgery. The goal is to fill all the black pixels that are inside the edge of the image into white pixels.

II. RELATED RESEARCH

There are several papers that discuss about segmentation and edge detection that is specifically on the medical part that detects objects in the image diagonasa as cancer of breast cancer, prostate, skin and others.
For example to develop the results of research and papers by previous authors who became the development materials as follows:
Image Magnetic Resonance Image (IMR) Imaging Using Edge Detection For Breast Cancer Identification [1] ▪ .
In this research, F. Ervina, and K. Bambang performed the processing of Magnetic Resonance Image (IMR) image through segmentation and edge detection process, to assist doctors or radiologists in identifying breast cancer [1] ▪ .
Identification of Prostate Cell Cancer Using Segmentation Methods Based on Object Size in Image [2] ▪ .
In this study Witeti identification of prostate cancer cells contained in digital images automatically by the method of segmentation based on the size of the object image pad [2] ▪ .A Novel Approach for Breast Cancer Detection and Segmentation in a Mammogram[3].
In this study K. Anuj and G.Bhupendra have improved the results of segmentation and edge detection with the development of methods to detect cancer areas and segment the areas covered by malignant tissue identified as cancer [3].
In this paper they focus on the malignant tissues that represent a higher intensity in comparison with the background.

III. METHODOLOGY

3.1 Computing design

In the whole stage of computing process in this study can be seen in the picture below.

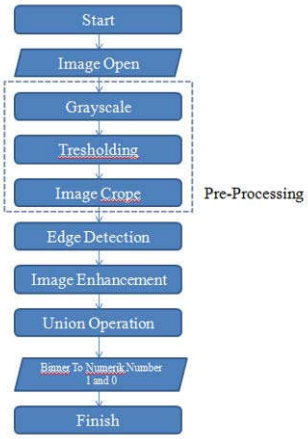


Fig. 1. Computing design process.

3.2 Stage of image processing

Pre-processing is a process used to improve the quality of the image display to have a better format so that the image becomes more easily processed further to produce the best edge detection. The stages of the pre-processing stage carried out in this study consisted of

1.Grayscale

Grayscaling is the process of converting an image that has a colour to an image that has a gray colour level. This process is done by converting the pixel value from 3 RGB values to 1 value. Although the results are quite good sometimes the use of the average value has not been optimal to show the grayscale image sometimes we can make the composition change so that the optimal results the author uses the equation.

$$S = 0.42r + 0.32g + 0.28b \quad (1)$$

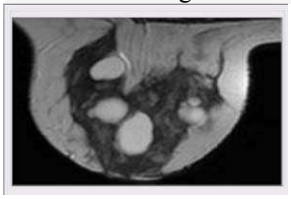


Fig. 2. Grayscale results

2. Tresholding

Tresholding or mining makes the image has tow gray levels of black and white, the mining process will produce binary image (7). The tresholding process follows the following rules.

$$f(x,y) = \begin{cases} 1 \text{ if } f(x,y) \geq T \\ 0 \text{ if } f(x,y) < T \end{cases} \quad (2)$$

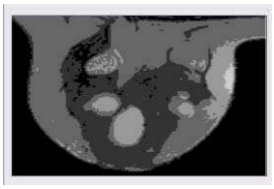


Fig. 3. Result of tresholding with threshold value (T=5)

3. Croping

Croping is intended to remove or eliminate unnecessary information. In the cropping process can be done after or before grayscaling and tresholding process.

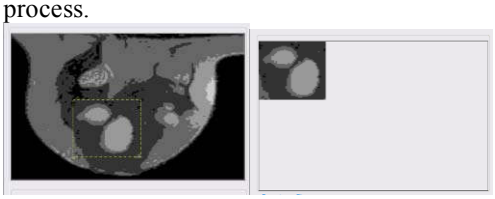


Fig. 4. Result of crop possible image objects identified

4. Edge Detection

Edge detection is a change of intensity value of a sudden large gray within a short distance [8]. edge characterizes the boundaries of the object and therefore the edge is useful for the process of segmenting and identification of objects in the image. In this process the MRI image here is done with edge detection developed from the second derivative of Robert. The filter kernel used in the edge detection is used

$$h = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} v = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \quad (3)$$

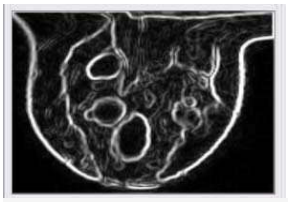


Fig. 5. Edge Detection Results

5. Image Enhancement

The image of edge detection that is still not maximal we can fix, in this peneliatian with approach of High Pass Filter, with aim to generate high frequency value so as to form more prominent edge.



Fig. 6. Results of Image Enhancement

6. Union Operation

Union is an amalgamation of two sets of purposes in use in this research that is the image forming process which initially only shows the white pixels at the edge of the result of the edge detection, then we fill all the images with white pixels, the purpose is to compare the number of binner values 0 and 1. The set notation for union operation can be written.

C = A U B (4)

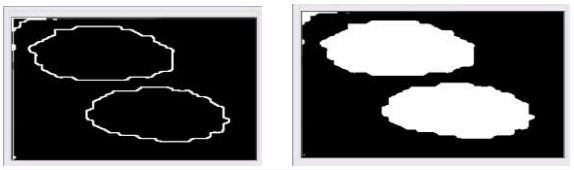


Fig. 7. Edge Detection Result Modified to Binary Image Union Operation

7. Binaryization

Binary process is a process that allows each pixel of an image to be expressed with a value of two possibilities. In this study, the binary algorithm is determined by the value of the image change to grayscale or called intensity, which if the intensity < 128 is automatically expressed by 0 and if not worth 255 or 1. As for the rules of the equation as below.

$$f(x,y) = \begin{cases} 1 \text{ if } f(x,y) \geq 128 \\ 0 \text{ if } f(x,y) < 128 \end{cases} \quad (5)$$

As for the conversion of binary images to numerical values 0 and 1

$$f(x,y) = \begin{cases} 1 \text{ if } r <> 255 \text{ and } g <> 255 \text{ and } b <> 255 \\ 0 \text{ if } r = 255 \text{ and } g = 255 \text{ and } b = 255 \end{cases} \quad (6)$$

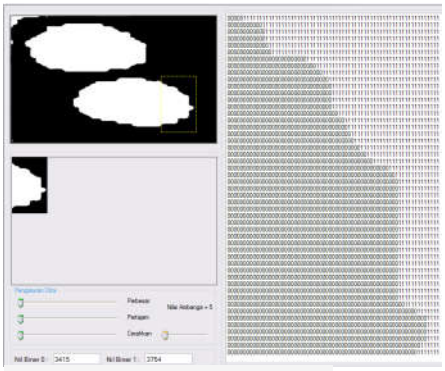


Fig. 8. Results Binary Process

8. Anastasia image to make decision

Analyze the ratio of the results of binaryisasis to conclude whether the image concerned is likely to suffer from breast cancer or not. The rules in use when the binary value 0 > binary value 1 with a very neat edge pattern and analysis of experts who have understood this field. It means that most patients diagnosed with breast cancer. If the binary value is 0 < binary value 1 with irregular edge pattern then vice versa.

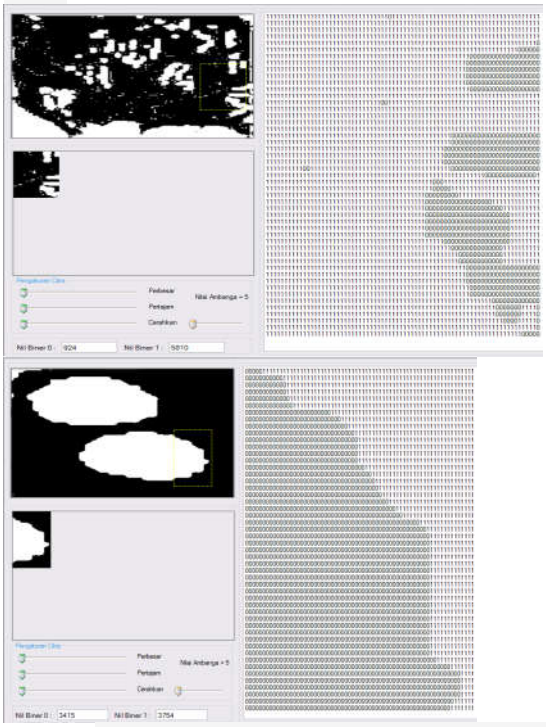
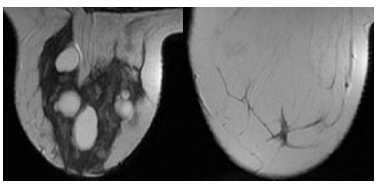


Fig. 9. Anilysis Binary Pattern And Comparison Of The Number Of Values 0 And Value 1

IV.RESULT AND DISCUSSION

4.1 Program Testing

The test is based on the quality of the resulting edges then converted into union process and tested by comparison of small values of 0 and 1 from binary with a regular pattern or not. In the table below the author uses 4 sample images 2 normal images and 2 abnormal images with the same image position. With the aim as a comparison material to be able to determine the final value of whether the patient contracted breast cancer or not.



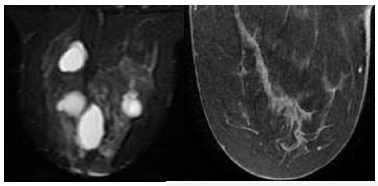


Fig. 10. Testing Image

1. Grayscale process

The grayscaling process is performed on all 4 test images, since the overall MRI image with high tesla has a depth of 16 bits, so it needs to be converted into an image with a depth of 8 bits, to be easily processed for subsequent processes. For example, following one of the grayscaling process image results for norm and abnormal images.

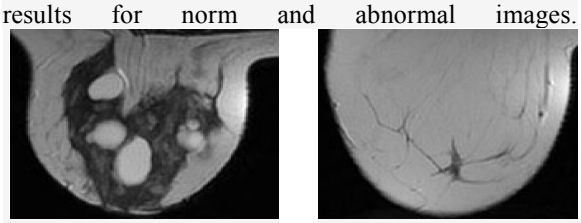


Fig. 11. Abormal Image and Normal Image

2. Tresholding

In this study the threshold value used for the entire test image is T = 5. The determination of this threshold value is based on the results of trials conducted with some normal and abnormal breast images, this is adjusted to the needs as a step before entering on binary. Therefore the determination of the threshold value must be appropriate and appropriate. The results of testing tresholding against some normal and abnormal images can be seen in the picture below 1 T = 5 and Figure 2 T = 100.

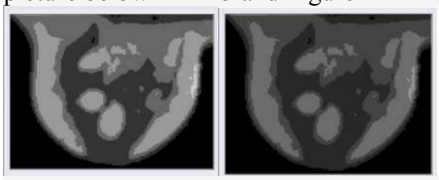


Fig. 12. tresholding results with threshold values of 5 and 100

3. Cropping

Before doing image edge detection we first croping, in this research croping process is done in addition to maintaining the rights and privacy of patients, patient data information contained in the image of trials obtained from the Family Hope Hospital needs to be removed. Also the size of the image used as test data is very large and varied, it affects the process of image processing, therefore the croping process is needed. As an example :

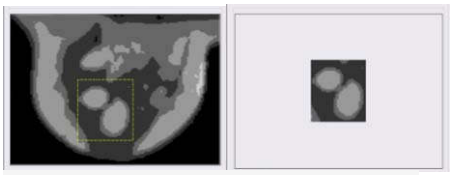
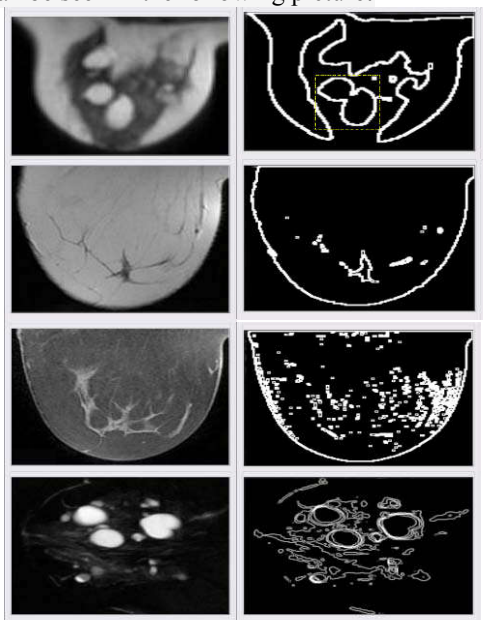


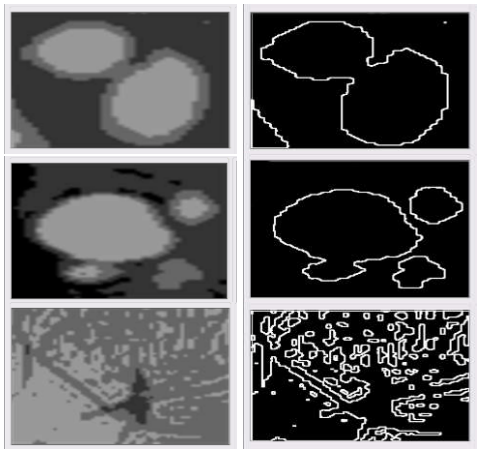
Fig. 13. Croping And Croping Results

4.2 Edge Detection Testing

After we do crope we will do edge detection. The edge detection process using the Prewitt operator of the 4 sample imagery samples consisting of 2 abnormal images and 2 normal images produces an image with borderline sightings of each region or clearly visible objects and no discontinuous edges. As for one of the results of edge detection testing of some image of preprocessing result can be seen in the following picture:



The preprocessing image that has been in crope first



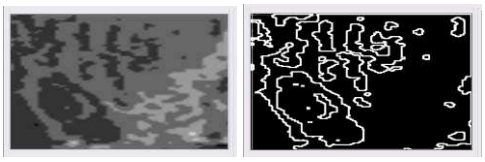


Fig. 14. Process Edge Detection Image large and crope result

In the crope image, it is the same part of abnormal object and normal object. We can see the abnormal edges more regularly while the normal mess.

4.3 Testing Binary and Union Operations

In converting the image into binary image form in this study using the method that is by binerisasi and tresholding [9]. If in tresholding the determination of threshold value is done manually, ie by shifting the trackbar to give the threshold value if the intensity value < 128 then the pixel value is expressed with 0 and if it is not then 1 [10] [11]. However, in this study the authors prescribe the threshold value T = 5. In addition the value of 0 and 1 are generated after we perform union operations from the image results from edge detection is displayed in a memo text in order to calculate the total occurrence of values 0 and 1 partial reference to compare the magnitude of values 0 and 1.

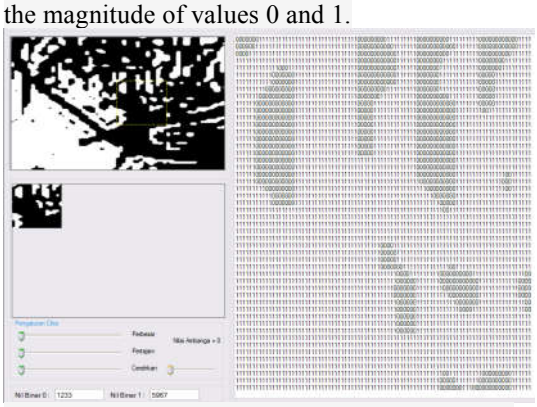


Fig. 15. Binary Process Image Crope Normal 1

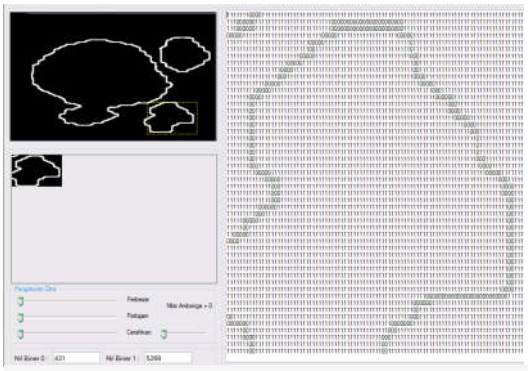


Fig. 16. Image Abnormal Image Crope Binaryization Process 1

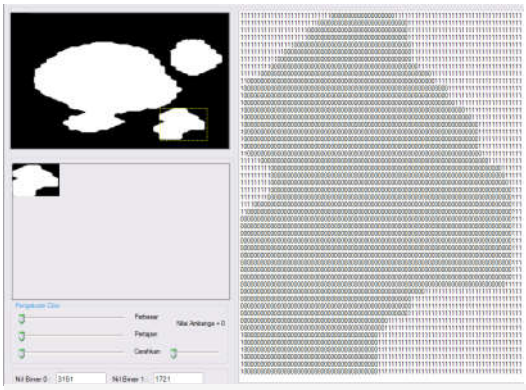


Fig. 17. Image Abnormal Image Crope Binaryization Process With Union Operation

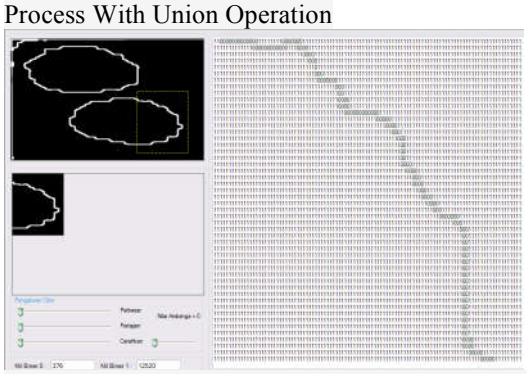


Fig. 18. Image Abnormal Image Crope Binaryization Process 2

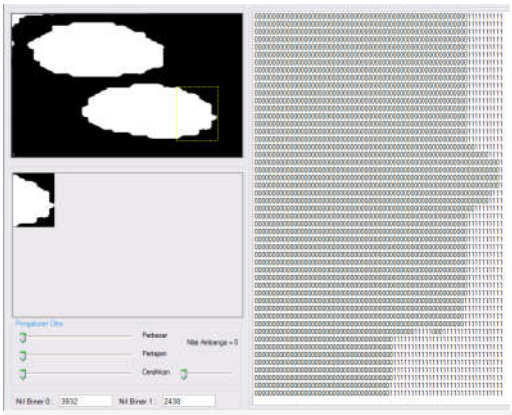


Fig. 19. Image Abnormal Image Crope Binerization Process 2 with Union

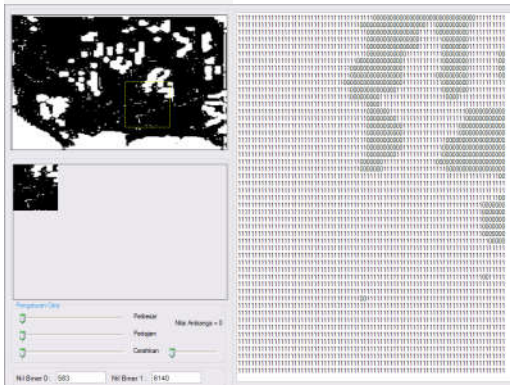
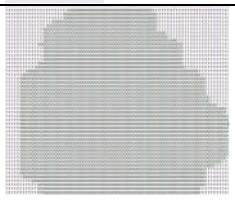


Fig. 20. Image Normal Image Crope Binerization Process 2

Finally we can see the comparison of the reference value in this case the author uses a sample of images that are interrelated between abnormal and normal images as in the table below.

Tabel I Final Score Comparision

Name	Crope Abnormal image	Crope Normal image
Image		
Binary		
Value	0 = 4168 1 = 2264	0 = 251 1 = 2609
Pattern	Regular	Irregular

The equations are in use

$$f(x,y) = \sum_{y=0}^{n} \begin{cases} if(r = 255 \text{ and } g = 255 \text{ and } b = 255)\, i++ \\ if(r <> 255 \text{ and } g <> 255 \text{ and } b <> 255\, j++) \end{cases} \quad (7)$$

Information:
i = Variable that holds the total value 0.
j = Variables that hold a total of 1.

V.CONCLUSION

1. To determine whether the patient is likely to develop breast cancer or can not see the comparison of binary values 0 and 1 if the value 0 > of value 1 with a regular edge pattern of puncture then the possibility of breast cancer and if the value 0 < of the value 1 with the edge determination pattern which is messy and irregular then it is likely not to suffer from breast cancer.
2. Image MRI when in though must pass through grayscale and tresholding first so that the edge detection process can determine the edge well, besides we can make edge detection improvements with shorter on High Pass Filter to improve the image to have a very good edge
3. Edge edge detection value 0 for white color and value 1 for black color without us using union image operation then automatically if calculated clearly value 0 will be smaller than 1 so we have to use union image operation to fill black color in circle edge and become all white, so the value 0 will be greater than value 1.
4. In the previous study only reached the segmentation stage and the conversion of binary images to numeric values of 1 and 0, while in this study there was an image improvement process after edge detection and the addition of union

operations to make numerical values more precise. And this value is also displayed. While previous research values are only displayed without calculating.

5. For the next development, it is expected to be able to do clusters of the binary image values that have been produced.

REFERENCE

[1] E. Varijki, B. T.-J. Matrik, and undefined 2017, "Segmentasi Citra Mri Menggunakan Deteksi Tepi Untuk Identifikasi Kanker Payudara," *Jurnal.Stmikbumigora.Ac.Id*.

[2] W. G. Flores and W. C. de A. Pereira, "A contrast enhancement method for improving the segmentation of breast lesions on ultrasonography," *Comput. Biol. Med.*, vol. 80, no. November 2016, pp. 14–23, 2017.

[3] B. Rami, A. Zlotnick, and S. Hashoul, "A WEAKLY LABELED APPROACH FOR BREAST TISSUE SEGMENTATION AND BREAST DENSITY ESTIMATION IN DIGITAL MAMMOGRAPHY Rami Ben-Ari , Aviad Zlotnick , Sharbell Hashoul IBM Research - Haifa , Israel," no. 4, pp. 722–725, 2016.

[4] R. Mehra and L. Sharma, "Comparative Analysis of Canny and Prewitt Edge Detection Techniques used in Image Processing," vol. 28, no. 1, pp. 48–53, 2015.

[5] M. A. Soeleman, M. Hariadi, and M. H. Purnomo, "Adaptive threshold for background subtraction in moving object detection using Fuzzy C-Means clustering," *TENCON 2012 IEEE Reg. 10 Conf.*, pp. 1–5, 2012.

[6] G. N. Chaple, R. D. Daruwala, and M. S. Gofane, "Comparisions of Robert, Prewitt, Sobel operator based edge detection methods for real time uses on FPGA," *Proc. - Int. Conf. Technol. Sustain. Dev. ICTSD 2015*, no. 1, pp. 4–7, 2015.

[7] S. N and V. S, "Image Segmentation By Using Thresholding Techniques For Medical Images," *Comput. Sci. Eng. An Int. J.*, vol. 6, no. 1, pp. 1–13, 2016.

[8] W. He, P. Hogg, A. Juette, E. R. E. Denton, and R. Zwiggelaar, "Breast image pre-processing for mammographic tissue segmentation," *Comput. Biol. Med.*, vol. 67, pp. 61–73, 2015.

[9] A. Makandar and B. Halalli, "Threshold Based Segmentation Technique for Mass Detection in Mammography," *J. Comput.*, vol. 11, no. 6, pp. 472–478, 2016.

[10] M. Holba, P. Bilík, and M. Kelnar, "Image processing in defectoscopy," *IFAC-PapersOnLine*, vol. 28, no. 4, pp. 65–70, 2015.

[11] S. Anitha and V. Radha, "Enhanced switching median filter for denoising in 2D patterned textile images," *Procedia Eng.*, vol. 38, pp. 3362–3372, 2012.

[12] Gonzalez, R. C., Woods, R.E., (2002). *DigitalImage Processing second edition*, New Jersey: Prentice-Hall, Inc.

Identification of Regional Dialects Using Mel Frequency Cepstral Coefficients (MFCCs) and Neural Network

Ayu Mawadda Warohma
Department of Electrical Engineering
Faculty of Engineering, Universitas Sriwijaya
Ogan Ilir 30662 South of Sumatera, Indonesia
ayumawaddawarohma@gmail.com

Puspa Kurniasari
Department of Electrical Engineering
Faculty of Engineering, Universitas Sriwijaya
Ogan Ilir 30662 South of Sumatera, Indonesia
puspakurniasari@gmail.com

Suci Dwijayanti
Department of Electrical Engineering
Faculty of Engineering, Universitas Sriwijaya
Ogan Ilir 30662 South of Sumatera, Indonesia
suci.dwijayanti@gmail.com

Irmawan
Department of Electrical Engineering
Faculty of Engineering, Universitas Sriwijaya
Ogan Ilir 30662 South of Sumatera, Indonesia
Irmawan@unsri.ac.id

Bhakti Yudho Suprapto
Department of Electrical Engineering
Faculty of Engineering, Universitas Sriwijaya
Ogan Ilir 30662 South of Sumatera, Indonesia
bhakti@ft.unsri.ac.id

Abstract*— Dialects may affect the voice recognition because they have an influence on the intonation and the pronunciation of a syllable. Every dialect shows the characteristic of a particular tribe and region. This paper addresses to develop a speech recognition system using Mel Frequency Cepstral Coefficients which are fed to a neural network based on backpropagation algorithm to recognize some particular dialects. The regional dialects used in this paper are Indonesian dialects, namely, Bataknese, Minang, and Javanese. The experiments are conducted to recognize the dialects from training and testing data. The testing data used in the experiments are from different people who have never been trained before so that it is expected that the recognition results will be more valid. The accuracy rate is used to evaluate the performance of the system. The accuracy of dialects recognition for Bataknese, Javanese and Minang are less than 20%, 50% - 80% and 70% - 90%, respectively. This method has succeeded in identifying the regional dialects and the most precise identification is on Minang and Javanese dialects.

Keywords : Backpropagation, Javanese dialect, Bataknese dialect, Minang dialect, MFCC, Neural Network.

I. INTRODUCTION

As an archipelagic state, Indonesia has a diversity of languages in each region. It is estimated that there are 700 variations of regional languages used in Indonesia. Most regions have their own local language which is very different from Bahasa Indonesia as an official language and natives tend to use the local language for daily conversation. Hence, the characteristic of local dialects could be heard when they speak Bahasa Indonesia. The dialect may indicate their origin or where they come from. The dialect is considered to be a major obstacle for voice recognition systems because a strong dialect may affect the intonation and pronunciation of syllables[1]. However, there are a few research on dialect identification in Indonesia so that it is still very possible to develop it in the future. Some studies on dialects have been shown in [2] that study of dialect on Bangladeshi, Persian accent[3], and the Tunisian dialect[4]. Rahmawati and Lestari [1] studied local dialects in Indonesia, Javanese and Sundanese. Some methods have been used to recognize such dialects. The spectral features and prosodic features were used as the input of Neural Network for Hindi dialect detection[5], Mel Frequency Cepstral Coefficients (MFCCs) and Support Vector Machine (SVM) were utilized for Pashto dialect[6], Gaussian Mixture Models (GMM) and MFCCs were for Bangladeshi dialect [2]. Results from various research have shown that MFCCs are reliable features to recognize the voice with the accuracy over 80%.

In this paper, the MFCCs are used in a feature extraction stage because of their sensitivity and capability to capture the main characteristic of the speaker's voice, the high accuracy of the voice recognition, and low complexity. A Backpropagation Neural Network algorithm (BPNN) is then utilized to recognize the dialects. The BPNN has adaptive learning capabilities for the training data provided and generalized to new conditions [1]. In BPNN, some statistical parameters are also altered to further enhance the accurate detection of the spoken words. Changing the number of hidden neuron units in each hidden layer of BPNNmay affect the recognition accuracy of the three dialects studied in the research.

Here, three dialects, namely Javanese, Bataknese and Minang, are used to be identified. These dialects are chosen because they have unique characteristics and the speakers may carry them while speaking Bahasa Indonesia. Dialect of Bataknese has a characteristic of a loud and high intonation, and t some emphasis after each sentence. While the Javanese dialect has a tendency to add the sound of [h] and silent sound, and the pronunciation of /b/j/d/ seems like spurted commonly known as “medok”. Then, dialect of Minang is very typical with the pronunciation of the letter /e/ which is thick and spoken in fast tempo. Each of these dialects has a distinctive intonation when they are spoken in some words.

Following this section, the discussion of the paper is divided into several parts. Section 2 explains the theories and

methods used in this study. In Section 3, we describe the design and implementation of the proposed method. Section 4 provides the experimental results and the discussion of the paper. In Section 5, the conclusion is presented.

II. Modelling of Dialect Identification

In the initial process of local dialect identification, silence is removed from the speech. Here, we use a speech corpus which are in the form of sentences spoken by different speakers. Those sentences have the same duration of 3 seconds for each speech sample. A feature extraction is conducted to obtain MFCCs with 13 coefficients. We consider utilizing MFCCs because of their ability to compactly represent the speech amplitude spectrum. These MFCCs are then fed to BPNN to recognize the dialects.

A. Silence Removal

The dialect recognition system generally begins with silence removal [3]. In this study, short time energy method is used to determine speech and non-speech parts of the signals. Only the speech parts are then considered to be processed to obtain MFCCs. Frame energy used in this step is 0.01 from the average energy of all waveforms.

B. Feature Extraction using MFCC

Feature Extraction is obtained by several steps as shown in Figure 1.

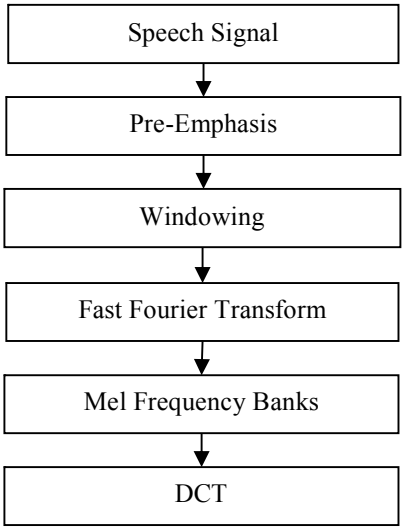


Fig. 1 MFCC Flow Chart

1) Pre-Emphasis

Pre-emphasis aims to allow the signal spectrum to be evenly distributed across all frequencies. It removes noise and maximizes its energy. In this process, the filter suppresses the lower frequency while leaving the higher frequency. Pre-emphasis filter is calculated as follows

$$H(Z) = 1 - \alpha z^{-1}, 0.95 \leq \alpha \leq 1 \quad (1)$$

Here, we use α of 0.97. The pre-emphasis output with the filter model as in (1) can be written as [6]

$$g(n) - x(n) - \alpha(n-1) \quad (2)$$

where *x(n)* is symmetrical window function.

2) Framming

The audio signals change over time and are generally non-linear which make the analysis be difficult. Thus, framing is needed to break down the digital signal into groups of a certain time which has smaller dimensions than the previous signal. To analyze it, the audio signal is divided into short frames with the lengths varying between 20-40ms. Frame size should not be less than 20 ms because it may cause each frame to lose its spectral estimated reliability and its differentiating power may decrease. In this study, the standard frame size for each signal is 25 ms and 16 kHz per second is taken as the sampling frequency. Thus, it has a 60% overlapping window between the frame and audio signal. The sound signal is then blocked into the frame with N sample. The adjacent frames are separated as far as M sample. Each frame has N sample and separated as far as M sample. If M << N, the spectral prediction of each frame becomes better.

3) Windowing

Windowing is utilized to minimize the effect of signal discontinuity at each beginning and end of the frame. It divides the signal into small frames so the analysis in the time domain can be calculated easily. It also reduces the signal to zero at the beginning and end of each frame. The window used for this study is Hamming Window as follows [6]

$$W(n) = \left(0.54 - 0.46\cos\left[\frac{2\pi n}{N-1}\right]\right) \quad (3)$$

N represents the total number of points in the window and n is an integer.

4) Fourier Transform

The next step is to perform a Fourier transform on each sample of the sound signal to measure the response of the signal magnitude. It is necessary to convert the audio signal from the time domain into the frequency domain. Short time Fourier transform can be calculated as follows [6]

$$X_1(k) = \sum_{n=0}^{N-1} x_n e^{\frac{j2\pi nk}{N}}, \qquad k = 0,1,\ldots,N-1 \quad (4)$$

k represents frequency variable and *N* represent number of discrete time sample

5) Mel Frequency Banks

Frequency resolution adjustment is then performed by a perceptual frequency scale that meets the human ear characteristics in accordance with Mel stage filter bank. However, the frequency limits which can be achieved by the FFT spectrum do not correspond to the linear scale from the phonetic signal. Therefore, the triangle filter bank, which is a filter that has the frequency of response at each triangle-shaped height, is used to overcome this problem [6]. . It is necessary to be used to calculate the values of component weights of the spectral filter so that its response can be converted to Mel scale as in the following equation [6]

$$mel = 2595 * log_{10}\left(1 + \frac{f}{700}\right), \quad (5)$$

where *f* denotes the signal frequency.

6) Discrete Cosine Transform (DCT)

DCT is performed to convert the frequency domain back into the original time domain so that the analysis can be carried out on it. The following equation is to perform the DCT [6]

$$c(n) = \sum_{m=1}^{M} Y(m) \cos\left[\frac{\pi n(m-\frac{1}{2})}{M}\right] \quad (6)$$

where *C(n)* represents the MFCC, *m* is the number of the coefficients , *N* is the number of triangular bandpass filters, M is the total number of mel-scale cepstral coefficients and *Y(m)* results from spectrum multiplication with conjugate.

Based on the calculation results using DCT, 13 coefficients are obtained from the process of MFCCs characteristic extraction and they are used as input to Neural Network

C. Backpropagation Neural Network Algorithm

A Neural Network has a learning system that is able to recognize a system based on its characteristics. With such learning system, the Neural Network acquires knowledge even though there are disturbance and uncertainty. It can also represent knowledge flexibly by creating self-representation through self-regulation or learning ability (self-organizing). In addition, it is also able to learn from the tolerance of an error and process the knowledge efficiently with its parallel system [3][7].

In this paper, backpropagation which is the most widely used and stable algorithm, is used as the learning algorithm to \overcome problems. The structure of Neural Network can be seen in Figure 2. The structure of the Neural Network consists of input layers, hidden layers and output layers. Here, the number of hidden layers are two and the number of neurons in each layer are altered to find the best learning outcomes.

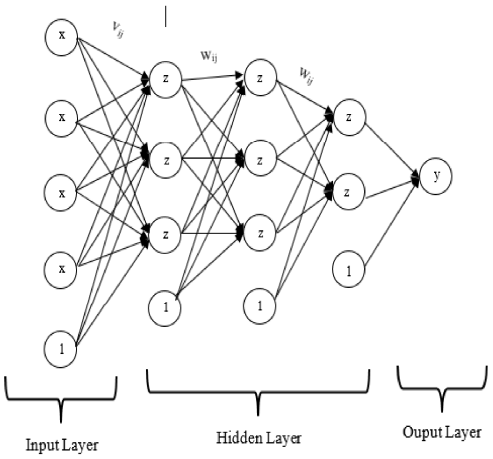


Fig .2 Neural Networks Architecture

III. Design and Implementation

The architecture design for the recognition of dialects being studied is shown in Figure 3. A training process is performed by the neural network followed by a testing process.

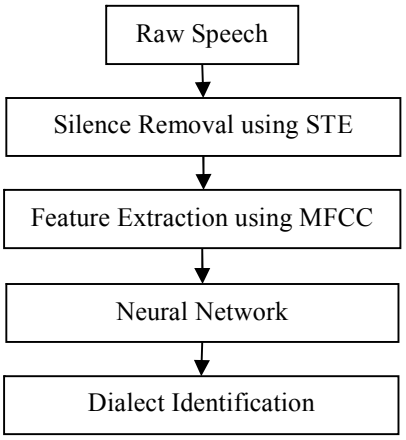


Fig. 3 Block Diagram of Dialect Identification Process

A. Database

Recording was done to obtain the speech signals which represent the dialects of Javanese, Minang, and Bataknese. Some local people with a strong local dialect were chosen as the representation of dialects. From this recording, we obtain database of dialects used in this study. A speech was recorded in a soundproof room and the sampling frequency is of 16 kHz. Since the objective of this study is to identify the regional dialect, all speakers speak the same clause in Bahasa Indonesia. It should be noted that some speakers who move from their place of origin or who do not speak their native language regularly may highly influence their dialect. Therefore, the selection of speakers affects the success rate of the dialect recognition significantly.

From the recording database, we have 120 voice samples which were divided into 40 voice samples of training data for each dialect area. Each dialect consists of 2 women and 2 men

who speak the specific clause and repeat it 10 times. The test data are obtained from 2 different people for each dialect since this study also measures the level of accuracy in recognizing the dialects coming from people who are not in the training data.. All selected speakers are at the age of 20-25. In total, there were 180 data which consist of 120 training data and 60 testing data.

B. Feature Extraction

The spectral features used in this study are MFCCs which are fed to the neural network. 13 coefficients of MFCC obtained by reducing the speech signal frequency information into values that mimic the vocal cords in the ear basilar membrane. To obtain these features, the signal is divided into short frames of 25 ms with 10 ms of frame shift and 20 filters bank are used to collect energy information from each frequency band.

C. Parameter of Neural Netwok

The process of classification and the recognition of dialect proposed in this study use backpropagation neural network with 3 hidden layers. These layers are sufficient to handle complex pattern of classification problem. The number of neurons in each hidden layers are altered and the testing process is conducted 4 times with different number of hidden units. In the hidden layer, a log sigmoid activation function is utilized. The function is most commonly used in the classification of a pattern and has range of 0 and 1. The learning rate is 0.01, a large learning rate may affect the system performance and too low learning rate may cause the long duration of the learning process.

D. System Evaluation

To evaluate the performance in identifying local dialects, an accuracy rate is measured. The following standard formula is used to calculate the accuracy ,

$$Accuracy = \frac{\text{number of data known}}{\text{total number of data}} x100\% \quad (7)$$

Besides the accuracy rate, mean square error (MSE) is also utilized to show the performance of the system.

IV. Experiment Results

MFCCs which are discussed in the previous section are used as the input to the neural network. The neural network us trained by using the BPNN algorithm with the obtained training data. Testing process is then performed to generalize the network by using the test data which are different from the training data.

In this study, four trials (tests) are conducted to evaluate the neural network performance. They were performed by changing the number of neurons in the hidden layer. Meanwhile, the number of input neurons was 13 neurons and the output was 1 neuron. The accuracy rate of each test is shown in Table 1. Table 2 shows the MSE for each test.

TABLE I. Performance of Dialect Recognition Accuracy

No	Dialek	Data	Accuracy			
			Testing 1	Testing 2	Testing 3	Testing 4
1	Bataknese	1	4.2%	6.9%	4.7%	10.5%
		2	8.0%	11.7%	14.3%	15.5%
		3	**8.6%**	**10.1%**	**12.1%**	**16.3%**
2	Javanese	1	52.6%	69.7%	73.4%	79.9%
		2	77.6%	82.4%	83.3%	83.4%
		3	**77.5%**	**70.3%**	**71.9%**	**76.7%**
3	Minang	1	71.2%	73.1%	77.4%	77.0%
		2	89.9%	93.0%	92.9%	93.2%
		3	**79.9%**	**81.4%**	**85.2%**	**83.9%**

This study consists of three experimental data for each dialect. Where the difference between the three experimental data is in the sample of voice used. The first to third data comes from different respondents' voice.

The first test was conducted by using 3 hidden layers with 7 neurons of first hidden layer, 4 neurons of the second hidden layer, and 7 neurons of the third hidden layer. The percentage of recognition accuracy in the first test is low. It may be caused by the number of hidden layer neurons which are small so the network is not able to resolve the problem of pattern recognition. The second trail, the number of neurons on the hidden layer are increased to 25 neurons for the first hidden layer, 15 neurons for the second hidden layer and 10 neurons for the third hidden layer. Thus, the percentage of accuracy improved. Then, in the third trial, the number of neurons in the hidden layer was changed into 26 neurons for the first hidden layer, 52 neurons for the second hidden layer, and 104 neurons on the third hidden layer. As the result, the percentage of the accuracy decreased in some test data. The dialect of Minang can be well recognized in this trial, but Javanese and Bataknese dialects were not yet well recognized. In the last trial, the numbers of neurons in the hidden layer were increased to 50 neurons for the first hidden layer, 75 neurons for the second hidden layer, and 100 neurons for the third hidden layer. In the fourth trial, the level of dialect accuracy increased significantly and an optimum value was obtained in which each dialect can be recognized. These results indicate that increasing the number of hidden layers may affect the accuracy of dialect recognition because the computing process becomes more complex. Thus, this experiment is done by using the number of neurons on the hidden layer as performed in the fourth test.

Table 2 shows the mean squares error (MSE) values of each trials. In the training process, we used 30,000 epochs. This training stops when the value of MSE was convergent and no significant change. The number of neurons in the hidden layer affects the time of the training process. The addition of neurons in the hidden layers make the learning process will be more thoroughly so that the number of error is

getting smaller as indicated by the value of MSE at each test. The MSE value is propotional to the level of recognition accuracy.

TABLE II. MSE of Dialect Recognition

No	Dialek	Data	MSE			
			Testing 1	Testing 2	Testing 3	Testing 4
1	Bataknese	1	0.399	0.282	0.251	0.224
		2	0.394	0.293	0.226	0.188
		3	0.442	0.292	0.229	0.214
2	Javanese	1	0.406	0.292	0.238	0.214
		2	0.398	0.293	0.227	0.193
		3	0.418	0.275	0.241	0.200
3	Minang	1	0.396	0.300	0.236	0.195
		2	0.399	0.307	0.222	0.188
		3	0.390	0.290	0.251	0.197

Figure 4 shows the performance of the regional dialect recognition expressed as the percentage of the accuracy level.

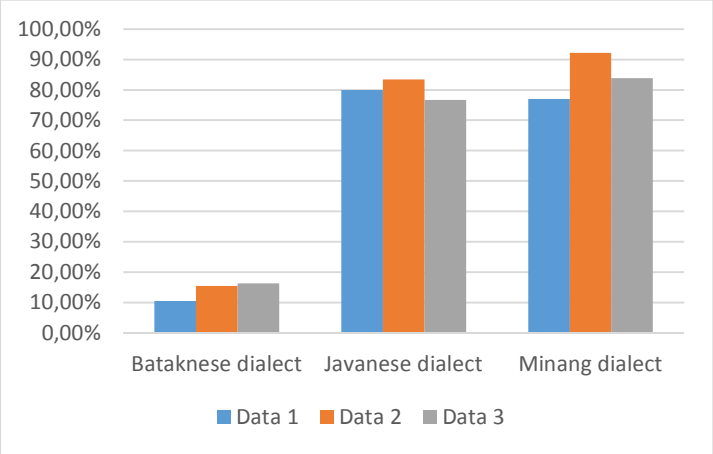


Fig.4 The Comparison of Recognition Accuracy Level

In the Fig. 4, the percentage of Bataknese dialect recognition for the first to the third data is in the range of 4% - 16%. This range of percentage is considered to be very low, it seems the BPNN gets difficulty to recognize this dialect well due to the level of Bataknese dialect variation and the difference of variation between the testing data and training data. It might be caused by some factors such as the dialect of Bataknese in the urban areas is different from the rural areas called as Karo Bataknese and Nias Bataknese dialects. For this reason, we need to consider a specification or uniformity between the test data and training data in the recognition of Bataknese dialect for a particular area.

For the Javanese dialect, the recognition percentage varies by 50% to 83%. As the case of Bataknese dialect, the Javanese dialect is also varied in each regions of Java so that the variation of the recognition becomes different. However, the Javanese dialect has a characteristic called "*medok*" or "*ngapak*" of which pronunciation in several letters such as /d/, /j/, and /b/ is stressed strongly. This typical characteristic makes the algorithm be able to recognize the Javanese dialect although it has variations.

The percentage of Minang dialect is not significantly different from the accuracy percentage of 70% - 90%. The fundamental difference of Minang dialect is in emphasizing intonation on letters like /e/ and spoken in a fast tempo. This characteristic may be recognized well by the BPNN algorithm as Minang dialect.

V. CONCLUSIONS

This paper discusses the dialects identification using MFCC and Neural Network. The experiments performed to three dialects in Indonesia, namely, Javanese, Minang and Bataknese dialects. The results show the sequence of the dialect recognition percentage. Javanese and Minang have higher accuracy level than Bataknese. It might be because Javanese and Minang have particular characteristics even though they are varied in each regions. Thus study shows that the characteristic of the dialects may affect the recognition. The success rate of recognition is also affected by the number of neuron. The more complicated recognition model, the bigger number of neuron is needed.

Further study can be explored to compare this proposed method with other recognition methods so that the advantages of each method can be highlighted. In addition, the recognition of a more specific dialect in one region that has a diverse dialect can also be studied. A dialect recognition system that involves gender as the factor of recognition can also be developed as Indonesia has ethnical, linguistic and cultural diversity

REFERENCES

[1] R. Rahmawati ; D. P. Lestari, "Java and Sunda Dialect Recognition from Indonesian Speech using GMM and I-Vector," in *2017 11th International Conference on Telecommunication Systems Services and Applications (TSSA)* , 2017.

[2] P. P. Das, S. M. Allayear, R. Amin, and Z. Rahman, "Bangladeshi dialect recognition using Mel Frequency Cepstral Coefficient, and Gaussian Mixture Model," in *Proceedings of the 8th International Conference on Advanced Computational Intelligence, ICACI 2016*, 2016, pp. 359–364.

[3] A. Rabiee and S. Setayeshi, "Persian Accents Identification Using an Adaptive Neural Network," *Educ. Technol. Comput. Sci. ETCS 2010 Second Int. Work.*, vol. 1, pp. 7–10, 2010.

[4] M. Hassine, L. Boussaid, and H. Massaoud, "Tunisian dialect recognition based on hybrid techniques," *Int. Arab J. Inf. Technol.*, vol. 15, no. 1, pp. 58–65, 2018.

[5] S. S. A. S. Sinha, A. Jain, "Speech Processing for Hindi Dialect Recognition," in *Advances in Signal Processing and Intelligent Recognition Systems*, pp. 161–169.

[6] S. Khan, H. Ali, and K. Ullah, "Pashto language dialect recognition using mel frequency cepstral coefficient and support vector machines," in *ICIEECT 2017 - International Conference on Innovations in Electrical Engineering and Computational Technologies 2017, Proceedings*, 2017.

[7] J. Gonzalez-Dominguez, I. Lopez-Moreno, P. J. Moreno, and J. Gonzalez-Rodriguez, "Frame-by-frame language identification in short utterances using deep neural networks," *Neural Networks*, vol. 64, pp. 49–58, 2015

LoG based Watermarking over Region Magnitude using Chinese Remainder Theorem

Prajanto Wahyu Adi
Department of Informatics Engineering
Universitas Dian Nuswantoro (UDINUS)
Semarang, Indonesia
prajanto@dsn.dinus.ac.id

Pramudi Arsiwi
Department of Industrial Engineering
Universitas Dian Nuswantoro (UDINUS)
Semarang, Indonesia
pramudi.arsiwi@dsn.dinus.ac.id

***Abstract*— An embedding model is proposed to hide the watermark in region magnitudes of the image using the Laplacian of Gaussian filter that is developed from the two previous researches that using block based and edge based watermarking through CRT algorithms. The block based CRT can cause the significant degradation of image when it lies on the homogeneous area. Meanwhile, the preceding study of edge based CRT detector is generated the embedding location as same magnitude in binary representation. The watermark bits are embedding on any edges location. The edge based CRT is unable to determine the most significant area for the embedding. The proposed method embeds the watermark bits from the highest region magnitudes in descending order of number of watermark bits. This is aimed to improve the previous researches that used block and edge based CRT algorithms. The experiments show that this model has higher structural similarity than the block based CRT which has average similarity value of 0.9996 and 0.9985 respectively. It also improved the robustness against compression the higher average correlation value of 0.6709 than previous CRT scheme which results in 0.6681 and 0.6616. The proposed model has better image quality of watermarked image in term of SSIM measurement and increase the robustness of watermark image in term of compression.**

Keywords—watermarking; laplacian of gaussian; chinese remainder theorem

I. INTRODUCTION

This paper presents advanced study from the two previous researches that using block based and edge based watermarking through Chinese Remainder Theorem algorithm. Originally, CRT [1] is an algorithm that is embedding the bits of watermark on the block of pixels within the host image size that divided by the watermark image size. The algorithm has great performance against additive noise and fair performance against compression, but it can cause the significant degradation on the homogeneous area of image.

In subsequent research [2], canny edge detector is used to define the proper embedding location. It is able to detect the heterogeneous areas of image that is then used to insert the bits of watermark. This scheme has better image quality and able to maintain the robustness against several attacks. However, canny edge detector is generated embedding location as same magnitude in binary representation. The watermark bits are embedding on any edges location. The edge based CRT is unable to determine the most significant area for the embedding.

Another studies of watermarking on the transform domain are conducted using discrete cosine [3], [4], discrete wavelet [5]–[8], and integer wavelet transform [9]–[11]. They are using the transformed the host image coefficients for watermark embedding that result in high computational cost which is unworthy with the improvement. Generally, the transformed domain methods have more complexity than the spatial methods due to the conversion process which involved such complicated mathematical operations. In fact, watermarking algorithms requires low complexity process for the online content protection and authentication.

A new watermark embedding process on the spatial domain is proposed in this study using Laplacian of Gaussian (LoG) filter. The filter is implemented on the host image to generate region magnitudes. The magnitudes are presented in unsigned integer 8-bit that can be used to determine the most proper embedding locations. The embedding of watermark bits is done from the highest region magnitudes in descending order of watermark bits number. The proposed scenario has higher similarity of watermarked image compared to the previous schemes that used block based and edge based CRT. It also preserves the robustness of the watermark image under compression and additive noise.

II. LAPLACIAN OF GAUSSIAN

LoG filter is the most used blob detection method that based on the Gaussian followed by Laplacian. This method is able to detect the regions that differ in properties. In digital image it compared a region with its surrounding pixel and then presents the region magnitude in integer representation of 8-bits. The initial process of LoG is applying the Gaussian as follows:

$$G\sigma(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e\left(-\frac{x^2+y^2}{2\sigma^2}\right) \quad (1)$$

where x and y is the pixel location, π is the kernel size, and σ is the standard deviation. Then, implements Laplacian operator to the pixel value $I(x,y)$:

$$L(x,y) = \frac{\partial^2 I}{\partial x^2} + \frac{\partial^2 I}{\partial y^2} \quad (2)$$

III. THE CRT

The CRT or Chinese Remainder Theorem is moduli based algorithm which able to factoring large number operation into using the mutually coprime of the number [12]. It is capable to express the value from the remainder series if all of the moduli are relatively prime [13]. CRT is rapidly developed in field of watermarking [14]. It can be used for watermark embedding by changing the coprime value [1]. Let:

$$y \equiv r_j \left(\mathrm{mod}\, z_j\right) \tag{3}$$

where y is 6 least significant bit of image elemen and r_i is the j-th remainder which is:

$$y \equiv \left(\sum_{j=1}^{s} r_j \frac{b}{z_j} k_j \right) \mathrm{mod} b \tag{4}$$

where z_j are coprime values of s series of number and the modulo b:

$$b = z_1 \cdot z_2 \cdot \ldots \cdot z_s \tag{5}$$

and then determine k_j:

$$\left(k_j \frac{b}{z_j} \right) \equiv 1 \quad \left(\mathrm{mod}\, z_j\right) \tag{6}$$

In conventional CRT model, the y altered in accordance with the bits watermark. The coprime values z_1 and z_2 are used to calculate the remainders r_1 and r_2 as show in equation (4). These remainders are used as a branching condition in watermark embedding which in turn will causes the value adjusting of y as shown in (5) and (6). The previous researches are based on the block and edge CRT to embed the watermark bits results in neither suboptimum imperceptibility nor robustness. Thus, this paper is proposed the use of region magnitudes of LoG filter by (1) and (2). This scheme is aimed to determine the most heterogeneous area of the host image in accordance with the watermark bits number which is aimed to improve the imperceptibility.

IV. THE PROPOSED METHOD

The embedding is done in highest frequency region of image sub band. The region magnitudes are used to define the most heterogeneous area of image. Fig. 1 shows the proposed model process like this:

A. Embedding

Based on previous research, the coprime value of 6 and 11 from the factor of 66 are used to generate the remainders r_1 and r_2 which is closest and larger than the highest 6-bit integer of 63.

1) Apply the Laplacian of Gaussian to get the region frequency
2) Read the bits number of watermark image
3) Get the most high region magnitudes in accordance with watermark bits number

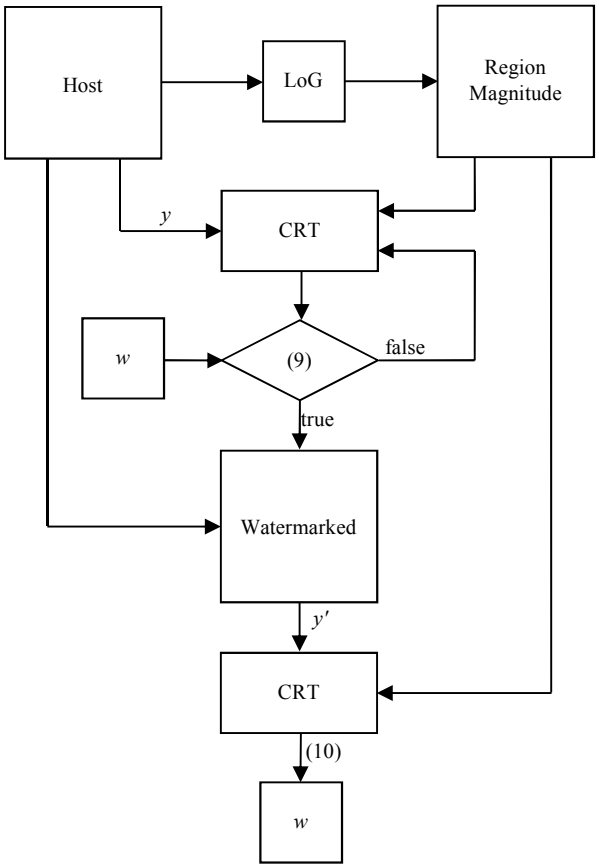


Fig. 1. Embedding and Extraction Model

4) Use the descending order of region magnitudes to define the most heterogeneous pixels
5) Read y
6) Get the remainders r_1 and r_2 using the coprime values

$$r_1 = y \bmod 6 \tag{7}$$

$$r_2 = y \bmod 11 \tag{8}$$

7) Adjust the y by the increment and decrement using a growth value for embedding the watermark bit w using:

$$r_1 \begin{cases} \geq r_2 & , w = 1 \\ < r_2 & , w = 0 \end{cases} \tag{9}$$

8) Iterate the step until the whole bits of watermark are embedded

B. Extraction

The highest region magnitudes are used to define the embedding location as follows:

1) Use the descending order of region magnitudes
2) Read the 6 least significant bit of y'
3) Generate the remainders r_1 and r_2 from the coprime values using (7) and (8)
4) Get the bits of watermark using the based on the remainders:

$$w = \begin{cases} 1 & , r_1 \geq r_2 \\ 0 & , r_1 < r_2 \end{cases} \quad (10)$$

5) Iterate the procedure until the entire bits of watermark are extracted.

V. Results and Discussions

To ensure stable assessment, this experiment is used the same images dataset from the previous research that compare [1] and [2] with the proposed method. Fig.2 shows the host images which consist of ten standard grayscale images with the dimension of 512x512 while the watermark image uses the UDINUS logo of 128x128. The watermark image is embedded to the host within 0 until 9 of seed value which start from the images of Baboon until Truck respectively. The seed is used to scramble the embedding location and generate the additive noise test that aimed to get stable results in robustness and imperceptibility. To give better representation the results are showed in table and chart.

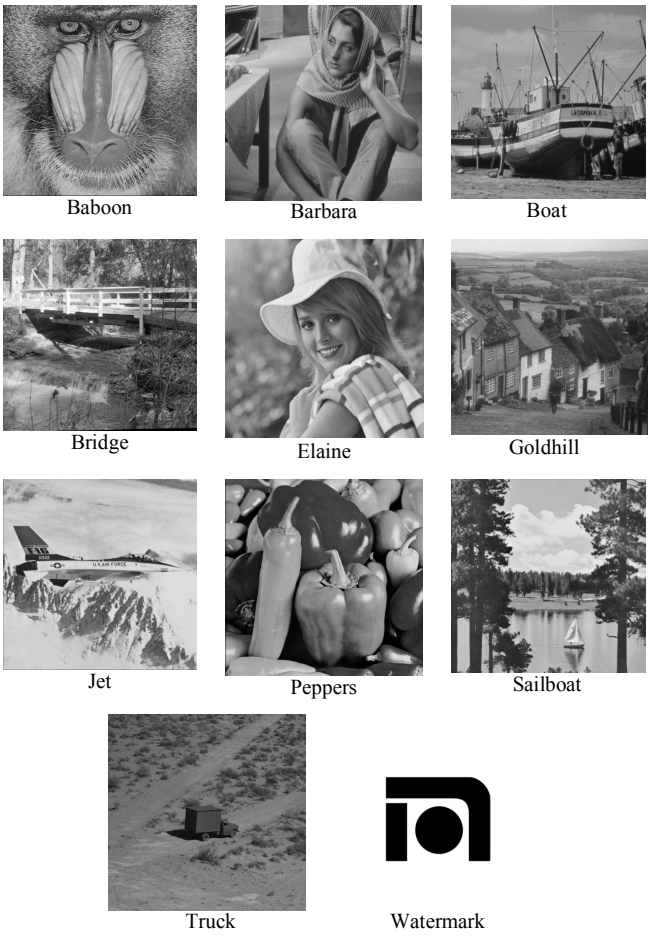


Fig. 2. The host and watermark images

A. Image Quality

The assessment of image quality are done using standard SSIM or Structural Similarity [15] is used in this experiment. It is able to represent the human visual system as a measurement standard.

TABLE I. Structural Similarity

Images	[1]	[2]	Proposed
Baboon	0.9994	0.9998	0.9999
Barbara	0.9985	0.9996	0.9999
Boat	0.9977	0.9994	0.9997
Bridge	0.9994	0.9997	0.9998
Elaine	0.9986	0.9991	0.9991
Goldhill	0.9986	0.9994	0.9995
Jet	0.9974	0.9995	0.9997
Peppers	0.9981	0.9994	0.9994
Sailboat	0.9988	0.9997	0.9997
Truck	0.9984	0.9992	0.9993
Average	**0.9985**	**0.9995**	**0.9996**

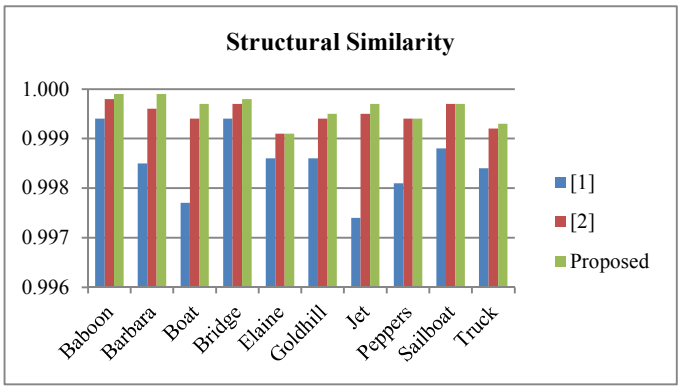


Fig. 3. Structural Similarity

Fig. 3 and Table I show that the model has better structural similarity of watermarked image than the conventional CRT model with the average values of 0.9996 and 0.9985 respectively.

B. Robustness under Compression

In this experiment, JPEG2000 with compression ratio of 3 is applied on the watermarked images. Meanwhile, to measure the correlation of watermarked images, the Normalized Correlation or NC is used.

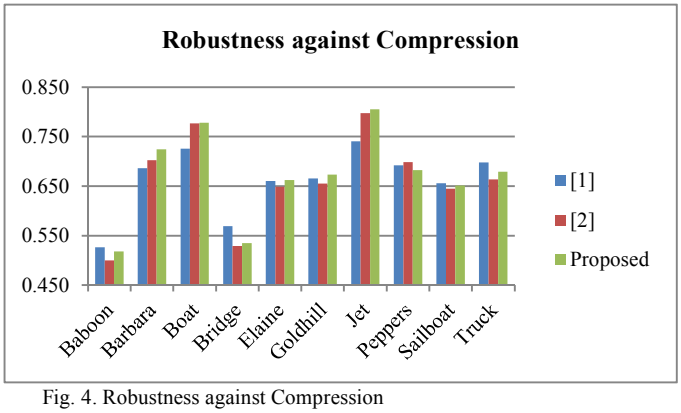


Fig. 4. Robustness against Compression

TABLE II. ROBUSTNESS AGAINST COMPRESSION

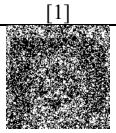
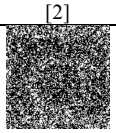
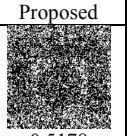
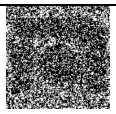
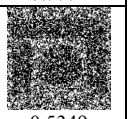
Images	[1]	[2]	Proposed
Baboon	 0.5264	 0.5001	 0.5178
Barbara	 0.6863	 0.7023	 0.7242
Boat	 0.7255	 0.7771	 0.7771
Bridge	 0.5689	 0.5290	 0.5349
Elaine	 0.6603	 0.6484	 0.6621
Goldhill	 0.6651	 0.6548	 0.6731
Jet	 0.7402	 0.7977	 0.8059
Peppers	 0.6921	 0.6985	 0.6823
Sailboat	 0.6555	 0.6450	 0.6512
Truck	 0.6980	 0.6634	 0.6791
Average	**0.6618**	**0.6616**	**0.6709**

TABLE III. ROBUSTNESS AGAINST ADDITIVE NOISE

Images	[1]	[2]	Proposed
Baboon	 0.9789	 0.9760	 0.9752
Barbara	 0.9738	 0.9742	 0.9759
Boat	 0.9741	 0.9749	 0.9732
Bridge	 0.9757	 0.9769	 0.9748
Elaine	 0.9743	 0.9721	 0.9770
Goldhill	 0.9759	 0.9762	 0.9748
Jet	 0.9734	 0.9768	 0.9777
Peppers	 0.9760	 0.9763	 0.9747
Sailboat	 0.9744	 0.9739	 0.9760
Truck	 0.9739	 0.9745	 0.9760
Average	**0.9750**	**0.9752**	**0.9755**

The proposed method is outperformed the previous methods in JPEG2000 compression test. The previous methods have the average NC values of 0.6618 and 0.6616 while the proposed method result in 0.6709 as shown in Fig. 4 and Table II.

C. Robustness under Noise

Last experiment is done using the 5% density 'salt and pepper' noise. The extracted watermark under the additive noise are compared to the original watermark to measure the correlation

The block and edge based are resulting in similar average value with the proposed method under the noise as shown in Table III and Fig. 5. The additive noise and compression test show that this model has better performance in term of robustness under noise and compression test. It also improves the imperceptibility. In the proposed model, the bits of watermark are inserted on the highest region magnitudes of transformed image. It has better structural similarity of watermarked image and preserves the watermark under JPEG2000 and 'salt and pepper' noise compared to the block and edge based CRT algorithms.

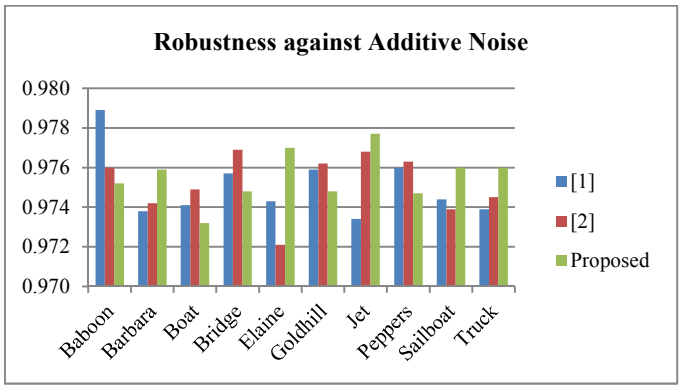


Fig. 5. Robustness against Additive Noise

VI. Conclusion

This paper proposed new watermarking model for embedding the watermark in region magnitudes of the image using the Laplacian of Gaussian filter. The most high region magnitudes are used to determine the most heterogeneous area for embedding the bits of watermark. The scheme is developed from the previous researches that used block and edge based CRT algorithms. Based on the results, the block based CRT has average structural similarity value of 0.9996 while the proposed method results in 9985. It also improved the robustness against compression the average NC value of 0.6709 in comparison with the block and edge based CRT which results in 0.6681 and 0.6616. Meanwhile, all of the methods have similar result in term of the robustness test under additive noise has. The proposed method has better structural similarity of the watermarked images. It also improves the correlation of extracted and original watermark images under JPEG2000 compression.

Acknowledgment

This research was funded by the Institute of Research and Community Services (LPPM) of Universitas Dian Nuswantoro through a Research Grants of 016/A.38.04/UDN-09/I/2018.

References

[1] J. C. Patra, A. Karthik, and C. Bornand, "A novel CRT-based watermarking technique for authentication of multimedia contents," *Digit. Signal Process.*, vol. 20, no. 2, pp. 442–453, 2010.

[2] P. W. Adi, Y. P. Astuti, and E. R. Subhiyakto, "Imperceptible Image Watermarking based on Chinese Remainder Theorem over the Edges," in *4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, 2017, pp. 1–5.

[3] J. C. Patra, J. E. Phua, and C. Bornand, "A novel DCT domain CRT-based watermarking scheme for image authentication surviving JPEG compression," *Digit. Signal Process.*, vol. 20, no. 6, pp. 1597–1611, 2010.

[4] C. A. Sari, D. R. I. M. Setiadi, and E. H. Rachmawanto, "Robust and Imperceptible Image Watermarking by DC Coefficients Using Singular Value Decomposition," *4th Int. Conf. Electr. Eng. Comput. Sci. Informatics (EECSI)*, no. September, pp. 19–21, 2017.

[5] C.-C. Lai and C.-C. Tsai, "Digital Image Watermarking Using Discrete Wavelet Transform and Singular Value Decomposition," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 11, pp. 3060–3063, 2010.

[6] C. Li, Z. Zhang, Y. Wang, B. Ma, and D. Huang, "Dither modulation of significant amplitude difference for wavelet based robust watermarking," *Neurocomputing*, vol. 166, pp. 404–415, 2015.

[7] M. Ali, C. W. Ahn, M. Pant, and P. Siarry, "An image watermarking scheme in wavelet domain with optimized compensation of singular value decomposition via artificial bee colony," *Inf. Sci. (Ny).*, vol. 301, pp. 44–60, 2015.

[8] P. W. Adi, Y. P. Astuti, and E. R. Subhiyakto, "Feature Image Watermarking Based on Bicubic Interpolation of Wavelet Coefficients Using CRT," *CommIT (Communication Inf. Technol. J. 11*, vol. 11, no. 2, pp. 93–99, 2017.

[9] N. M. Makbol and B. E. Khoo, "A new robust and secure digital image watermarking scheme based on the integer wavelet transform and singular value decomposition," *Digit. Signal Process.*, vol. 1, no. 134, pp. 1–14, 2014.

[10] V. Singh, R. Kumar, and A. Ojha, "Significant region based robust watermarking scheme in lifting wavelet transform domain," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 8184–8197, 2015.

[11] P. W. Adi, F. Z. Rahmanti, and L. Umaroh, "Improving CRT based watermarking using integer wavelet projection," in *2017 International Conference on Innovative and Creative Information Technology*, 2017, pp. 1–5.

[12] S. G. R. Ekodeck and R. Ndoundam, "PDF steganography based on Chinese Remainder Theorem," *J. Inf. Secur. Appl.*, vol. 29, pp. 1–15, 2016.

[13] X. Li, W. Wang, W. Wang, X. Ding, and Q. Yin, "Optimal estimates of common remainder for the robust Chinese Remainder Theorem," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 19, no. 7, pp. 2373–2381, 2014.

[14] U. Sudibyo, F. Eranisa, E. H. Rachmawanto, D. R. I. M. Setiadi, and C. A. Sari, "A Secure Image Watermarking using Chinese Remainder Theorem Based on Haar Wavelet Transform," *2017 4th Int. Conf. Inf. Technol. Comput. Electr. Eng.*, pp. 208–212, 2017.

[15] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

The Determination of Cluster Number at k-mean using Elbow Method and Purity Evaluation on Headline News

Dhendra Marutho[1], Sunarna Hendra Handaka[2], Ekaprana Wijaya[3], Muljono[4]
[1,2,3,4]Computer Science Faculty of Dian Nuswantoro University, Semarang, Indonesia
[1]dhendra@valudata.net, [2]nana_fe@mail.unnes.ac.id, [3]ekapranawijaya@gmail.com, [4]muljono@dsn.dinus.ac.id

***Abstract*— Information is one of the most important thing in our lives, while humans is naturally impatient when searching for information from the internet. Users want to get the right answer instantaneously with minimal effort. News headlines can be used to categorize news types, as appropriate. The appropriate type of news can make it easier for us to choose the particular topic we want. Similarity in a title can be used to clustering news based on news title. From those reason this dataset research contain the title of online news site. TFIDF used as Document Preprocessing method, K-Means as clustering method, and elbow method used to optimize number of cluster. Purity method applied to evaluate news title clustering as internal evaluation. SSE (Sum Square Error) of each cluster are calculate and compared to optimize number of cluster in the elbow method, the result of those comparison evaluate using internal evaluation called purity, purity value is conformity between cluster and ideal cluster. From the calculation of elbow method, the most optimal number of cluster are 8 cluster, there is 0.228 point between 7cluster and 8 cluster SSE value so the elbow form are made. Purity evaluation method generates value 0.514 in the number of cluster are 8, this is the highest value and the one closest to one rather than the other number of cluster which mean the most ideal. The conclusion is the elbow method can be used to optimize number of cluster on K-Mean clustering method..**

Keywords—Clustering; K-Means;ElbowMethod;Purity;Tf-idf;

I. INTRODUCTION

Thousands and even millions of news generated by various news sites every day. Political news, economy, social culture, and entertainment are continuously published every time through the Internet, while humans are naturally impatient when searching for information from the Internet; users want to get the right answer instantaneously with minimum effort [1]. In many applications, the title is the first thing users notice. Although only a word or phrase, it can dramatically change the overall message of the content. A proper title is very important, therefore it should be descriptive, concise and correct in grammar [2].

News headlines can be used to categorize news type; appropriate news type can make it easier for users to choose the particular topic they wanted. Grouping of news by title can be done by looking at similarities in a title [3].

The three problems in the clustering process are: (i) determining the size of the similarity between different elements (ii) applying efficient algorithms to find the group of elements, which are most similar to the unsupervised way and (iii) obtaining descriptions that can characterize the elements of the cluster[4].

K-means algorithm is a reliable algorithm for clustering process. The headline data are clustered so that the appropriate classes are obtained. The problem that often arises is the determination of the number of clusters. The exact number of clusters will show the maximum resemblance in each class generated. The Elbow method can be a principle method to determine the exact value for the number of clusters. The purity value is used to evaluate the result of the method, purposed for producing the expected cluster number.

II. RELATED WORKS

K-Means is a popular and simple clustering technique, but the results are based on the chosen cluster center so that it can easily generate local optimization. Since the cluster center is randomly selected, the center may be poorly selected. A research conducted by Aditi Anand Shetkar et al. proposed a K-means ++ algorithm to solve this problem, by spreading the cluster center evenly. In K-means ++, the first cluster center is randomly selected then it looks for another point based on the exact possibility. K-Means ++ gives better results than K-Means [5].

Ahmad Izzuddin et al used Principal Component Analysis (PCA) to reduce the lecturer performance data before the k-mean algorithm was applied. It is proven to be effective to improve a model quality. The study used DBI to measure the validity of the cluster. The result showed that DBI, in case of combining PCA and K-mean is smaller than the conventional k-mean[6].

Ni Putu Eka Merliana et al used the Elbow method to determine the best number of clusters in the k-means algorithm. First, calculate the SSE value of the specified cluster. Secondly, when a drastic decrease in the SSE value and there is no significant change, then the current Elbow point is determined. That point is the optimal k value [7].

The cluster validity of k-mean and fuzzy c-mean algorithms is measured using purity and entropy. Entropy uses

external class information and cluster purity is measured by the reference to the labels called entropy. The lower entropy means better clustering. The entropy strengthens when the objects truth in the cluster is more diverse. The higher entropy means the worse cluster. The number of disorders found using entropy [8]. Satya Chaitanya Sripada et al mentioned that the higher the purity value indicates a good cluster. Entropy is an inverted measurement, the lower the entropy value the better the clustering result.

Budi Santoso et al used the Genetic Algorithm (GA) to optimize the cluster's initial center on the k-means algorithm. The k-means algorithm is applied to the lecturer performance data in a Faculty of Computer Science of Brawijaya University in 2016. The result showed that GA-K-means algorithm has higher cluster quality that achieved 2,74 % compared to the k-means algorithm without Genetic Algorithm (GA), where the quality of cluster obtained using the Silhouette Coefficient method [9].

Kamaljit Kaur et al compared K-means method with Median-Based K-means. The method was applied to the datasets of the UCI machine learning repository, in which the Median-based k-means was used to select the initial centroids. The result was to determine the initial centroid first which is better than random selection[10].

III. METHOD

A. *Proposed System*

Optimizing number of cluster of K-Mean method is the main purpose of this study. Elbow method are choose to determine the number of cluster while purity are perform to evaluate.

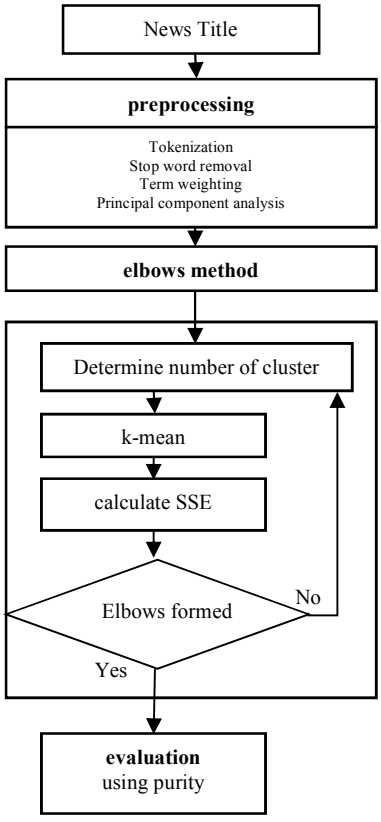


Figure 1. Design model of the proposed system

Here we got dataset contain news title from news site, the data are preprocessed using four type of preprocessing, first are tokenize next are stop word removal and TF-IDF term weighting are performed. The data reduced using principal component analysis.

K-Mean clustering method applied using number of cluster are 2 until 10, while SSE (Sum Square Error) are calculated and recorded. Put it on a graphic and determine the corner of the elbow method. Those result are evaluate using purity and each number of cluster are generate purity number.

B. *K-means*

K-means is a simple unsupervised learning algorithm, used to classify data based on Euclidian Distance technic between the data [11]. K-Means is a fast and simple clustering method with a smaller number of iterations. This algorithm divides data into *k* section. Cluster requirements are estimated based on user choice. Computers randomly select and assign objects to one cluster (k). The distance between each object and the center of each cluster is calculated and resulted in an optimal cluster solution. Objects within a particular cluster are adjacent to each other [12].

C. *Preprocessing*

1) Tokenize

Tokenization is the process of dividing a sentence into a word by omitting commas, spaces and special symbols in a sentence. This process generates tokens [5].

2) Stop word removal

Prepositions, articles, and pronouns etc, are the most common words in text documents and do not give any meaning to the document. These words are eliminated and not required for text mining applications [13].

3) Weighting

A weighting step is done by calculating the frequency of documents, one of the methods used is TF-IDF. After the pre-processing of the document, each document in the dataset is represented as an N-dimensional vector in the term space, where "N" denotes the number of words/terms. Document vectors are exposed to some standard weighting schemes, such as Term Frequency-Inverse Document Frequency (TF-IDF). We need to calculate the weight of Term Frequency (TF), Inverse Document Frequency (IDF) and finally TF * IDF i.e. TF and IDF products. TF-IDF analysis is done by considering two factors: Term Frequency (TF) and Inverse Document Frequency (IDF).

$$\text{TF-IDF} = \text{TF} * \text{IDF}$$

TF: - K / T, where K = the total number of certain words in document *d* and T = number of words in document *d*.

IDF = D / DF, where D = number of documents in dataset, DF = total number of documents containing a particular word [5].

D. *Principal Component Analysis (PCA)*

Principal Component Analysis (PCA) is a technique for connecting new variables that are linear combinations of the

original variable. The maximum number of these new variables will be equal to the sum of the old variables and are not correlated to each other [6]. The algorithm is used to reduce the data of several ratios into several indexes, which is a linear combination of all initial ratios [14].

E. *Elbow method*

This method focuses on the percentage of variants as the function of the number of clusters. Based on the idea that there should be an optimal number of k-means algorithm, so adding the number *k* will not contribute significantly [15]. The value of *k* is added one by one and the Sum Square Error (SSE) value is recorded.

$$SSE = \sum_{K-1}^{K} \sum_{xi \in Sk} \|Xi - Ck\|_2^2$$

SSE is the sum of the average Euclidean Distance of each point against the centroid [16]. When the value drops drastically and forms a smaller angle, then the value of *k* is found. Starting from k = 2 and the SSE value is then added step by step, where kn = k + 1, the largest SSE kn - SSE kn - 1 is the point in which the optimal *k* value is found. When the value of *k* is re-added then the new cluster is similar to the previous cluster or the number of errors does not change significantly which resulted in the value of *k*.

F. *Purity*

Purity is an algorithm used to evaluate the value of *k* produced by Elbow method. It is a method for measuring the quality of cluster based on the labeled data. The purity value 1 is gotten from the pure cluster result. That means that all data is perfectly classified according to the existing label. The resulting cluster is not pure if the purity value is close to 0 (zero), which means that multiple cluster members contained in different classes [17].

$$Purity = 1/N \sum_{i-1}^{k} max_j |C_i \cap t_j|$$

N = No. of data points
K = No. of cluster
C_i = Cluster in C
t_j is specify the max. count for each cluster C_i

IV. Result And Discussion

The dataset used contained a 1000 (thousand) news headline records retrieved from various news sites. The data is pre-processed i.e.

A. *Tokenization*

The tokenization process yields word data. From the headline dataset generated the words, which are used to the next steps.

B. *Stop word removal*

The next process is stop word removal, which removes prepositions, articles, pronouns etc.

C. *Term weighting*

Next step is the term weighting implemented to the data which then performed optimization using PCA technic so that it resulted in 500 data.

From the above processes, the data is then processed using the k-mean algorithm, starting with n = 2 to n = 10. In this process, the SSE value is revealed and showed in the following table.

TABLE I. K AND SSE VALUE

n	SSE	SSE ratio
2	0,877	0,877
3	0,720	0,157
4	0,698	0,022
5	0,677	0,021
6	0,659	0,018
7	0,655	0,004
8	0,427	0,228
9	0,419	0,008
10	0,422	0,003

From the above table, the largest difference is at *n* = 7 and *n* = 8 value, it is more clearly seen in Fig. 1 below

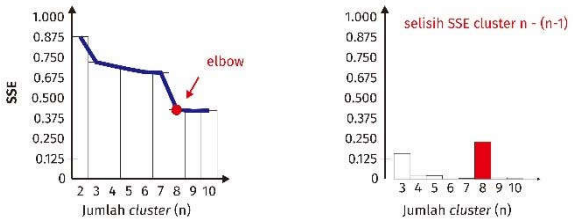


Figure 1. SSE value and Number of clusters

The value of SSE at n = 2 is 0.877 then at the value of n = 3 is 0.720, which create a difference of 0.157. The next is at n = 4 to n = 7 is 0.689, 0.677, 0.659, and 0.655. There was no significant change at these points, the different value ranges from 0.004 to 0.022. The change is quite drastic at the value of n = 8 i.e. SSE of 0.427, which means that there is a difference of 0.228, the difference is higher than 0.157. Then n value is increased successively to the 9 and 10 that yields SSE value of 0.419 and 0.422. That means there is a difference that is not too far which ranges in 0.008 and 0.003. It can be concluded that the *n* optimal point for the clustering data is n = 8 based on Elbow method.

The next step is to compare the result of the optimal cluster value using Elbow method and purity calculation for each value of *n*. The purity is the maximum number of each cluster divided by the amount of data. It used the class label as a category. In case of that, in this study the data processed amounted to 1000 with the label is as many as 4 pieces namely a, b, c, and d. The value of purity is the sum of the maximum value of each cluster divided by 1000. The value of n = 2 to n = 10 in arithmetic purity value, which are shown in the table below.

TABLE II. DATA *CLUSTER* N=2

Category	a	b	c	d
Cluster 1	0	0	66	0
Cluster 2	250	250	184	250

In the cluster number equal to 2, which most categories of data are inserted into cluster 2, only 66 from category *c* which becomes member of cluster 1. The maximum value at cluster 1 is 66 and on cluster 2 is 250 then obtained the purity value, which is equal to 0.316.

Furthermore, at the value of n = 3 to n = 7 that has a purity value ranging from 0.409 to 0.438, the value is still lower than the purity value of n = 8, is 0.514. The successive result of clustering n = 3 to n = 7 is in the following table.

TABLE III. DATA *CLUSTER* N=3

Category	a	b	c	d
Cluster 1	250	157	184	250
Cluster 2	0	0	66	0
Cluster 3	0	93	0	0

TABLE IV. DATA *CLUSTER* N=4

Category	a	b	c	d
Cluster 1	0	0	50	0
Cluster 2	250	157	171	250
Cluster 3	0	93	0	0
Cluster 4	0	0	29	0

TABLE V. DATA *CLUSTER* N=5

Category	a	b	c	d
Cluster 1	0	0	50	0
Cluster 2	250	146	171	250
Cluster 3	0	0	29	0
Cluster 4	0	39	0	0
Cluster 5	0	65	0	0

TABLE VI. DATA *CLUSTER* N=6

Category	a	b	c	d
Cluster 1	0	30	0	0
Cluster 2	0	0	50	0
Cluster 3	248	140	172	247
Cluster 4	0	0	28	0
Cluster 5	0	54	0	0
Cluster 6	2	26	0	3

TABLE VII. DATA *CLUSTER* N=7

Category	a	b	c	d
Cluster 1	0	30	0	0
Cluster 2	248	140	170	247
Cluster 3	0	0	40	0
Cluster 4	0	0	22	0
Cluster 5	0	0	18	0
Cluster 6	0	54	0	0
Cluster 7	2	26	0	3

The clustering value of n = 8 shown in Table VIII, produces a purity value of 0.514, which is a vertex on the Elbow method, based on that method the value at this point has to be the highest one. This is proven by the value of n = 9, in which the purity value decreased continuously into 0.503, shown in Table IX.

TABLE VIII. DATA *CLUSTER* N=8

Category	a	b	c	d
Cluster 1	0	29	0	0
Cluster 2	0	0	22	0
Cluster 3	0	0	40	0
Cluster 4	229	92	145	152
Cluster 5	0	0	18	0
Cluster 6	21	48	25	95
Cluster 7	0	29	0	3
Cluster 8	0	52	0	0

TABLE IX. DATA *CLUSTER* N=9

Category	a	b	c	d
Cluster 1	0	29	0	0
Cluster 2	0	29	0	2
Cluster 3	0	1	9	0
Cluster 4	0	0	39	0
Cluster 5	0	0	20	0
Cluster 6	0	0	18	0
Cluster 7	0	52	0	0
Cluster 8	20	39	22	77
Cluster 9	230	100	142	171

The purity value increased to 0.514 at the value of n = 10, the clustering results can be seen in Table X.

TABLE X. DATA *CLUSTER* N=10

Category	a	b	c	d
Cluster 1	0	21	0	0
Cluster 2	0	0	18	0
Cluster 3	0	32	0	0
Cluster 4	19	47	23	89
Cluster 5	229	92	141	158
Cluster 6	0	0	20	0
Cluster 7	0	1	9	0
Cluster 8	0	38	0	0
Cluster 9	0	0	39	0
Cluster 10	2	19	0	3

TABLE XI. *PURITY*

k value	*purity*
2	0.316
3	0.409
4	0.422
5	0.433
6	0.436
7	0.438
8	0.514
9	0.503
10	0.514

The clustering processes using k-means algorithm resulted in some groupings as shown in above tables. Table XI displayed purity calculation result on each specified n value. The maximum purity value is 0.514 which is generated in the value of n = 8 and n = 10.

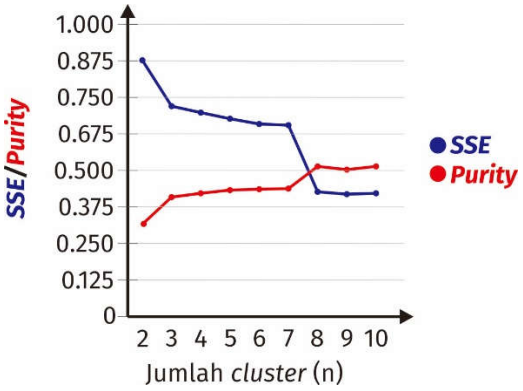


Figure 2. SSE and Purity

In Figure 2, the SSE and Purity values intersect at the value of n = 8 with the value of SSE = 0.427 and the maximum purity value is at 0.514.

V. CONCLUSION

the Elbow method is purposed for determining the number of the cluster in the k-mean algorithm that we adopt in this study. It is said that the optimal point for determining the *k* value is at the point, in which there is a significant change in the SSE value so that an angle is formed. The experiment was conducted on news headline data and performed an internal evaluation through purity. The purity test results showed that the number of the cluster in the Elbow method is same as the best result of internal purity evaluation measurement.

REFERENCE

[1] N. Gali, R. Mariescu-Istodor, and P. Fränti, "Using linguistic features to automatically extract web page title," *Expert Syst. Appl.*, vol. 79, pp. 296–312, 2017.

[2] H. Yunhua *et al.*, "Title Extraction from Bodies of HTML Documents and its Application to Web Page Retrieval," *Res. Dev. Inf. Retr.*, pp. 250–257, 2005.

[3] I. Blokh and V. Alexandrov, "News clustering based on similarity analysis," *Procedia Comput. Sci.*, vol. 122, pp. 715–719, 2017.

[4] A. Ahmad and L. Dey, "A k-mean clustering algorithm for mixed numeric and categorical data," *Data Knowl. Eng.*, vol. 63, no. 2, pp. 503–527, 2007.

[5] A. A. Shetkar and S. Fernandes, "Text Categorization of Documents using K-Means and K-Means ++ Clustering Algorithm," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 4, no. June, pp. 485–489, 2016.

[6] A. Izzuddin, "Optimasi Cluster pada Algoritma K-Means dengan Reduksi Dimensi Dataset Menggunakan Principal Component Analysis untuk Pemetaan Kinerja Dosen," *Energy J. Ilm. Ilmu-Ilmu Tek.*, vol. 5, no. 2, pp. 41–46, 2015.

[7] N. Putu, E. Merliana, P. Studi, M. Teknik, F. T.

Industri, and U. A. Jaya, "Analisa Penentuan Jumlah Cluster Terbaik Pada Metode K-Means," *Semin. Nasionalmulti Disiplin Ilmu&Call Pap. Unisbank*, pp. 978–979.

[8] S. C. Sripada, "Comparison of Purity and Entropy of K-Means Clustering and Fuzzy C Means Clustering," *Indian J. Comput. Sci. Eng.*, vol. 2, no. 3, pp. 343–346, 2011.

[9] B. Santoso, I. Cholissodin, and B. D. Setiawan, "Optimasi K-Means untuk Clustering Kinerja Akademik Dosen Menggunakan Algoritme Genetika," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 1, no. 12, pp. 1652–1659, 2017.

[10] K. Kaur, D. Singh Dhaliwal, and R. Kumar Vohra, "Statistically Refining the Initial Points for K-Means Clustering Algorithm," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 2, no. 11, pp. 2278–1323, 2013.

[11] S. S. Jamadar and P. D. Y. Loni, "Efficient Cluster Head Selection Method Based On K-means Algorithm to Maximize Energy of Wireless Sensor Networks," *Int. Res. J. Eng. Technol.*, vol. 3, no. 08| Aug 2016, pp. 1579–1583, 2016.

[12] M.Anusha, "An Enhanced K-Means Genetic Algorithms for Optimal Clustering," *Int. Conf. Comput. Intell. Comput. Res.*, vol. 14, pp. 1–13, 2016.

[13] M. Raghuvanshi and R. Patel, "An Improved Document Clustering with Multiview Point Similarity / Dissimilarity measures," *Int. J. Eng. Comput. Sci.*, vol. 6, no. 2, pp. 20285–20288, 2017.

[14] S. Sharma, "Applied Multivariate Techniques," *John Wiley Sons Inc.*, 1996.

[15] P. Bholowalia and A. Kumar, "EBK-Means : A Clustering Technique based on Elbow Method and K-Means in WSN," *Int. J. Comput. Appl.*, vol. 105, no. 9, pp. 17–24, 2014.

[16] E. Muningsih and A. B. S. I. Yogyakarta, "Optimasi jumlah cluster k-means dengan metode elbow untuk pemetaan pelanggan," *Pros. Semin. Nas. ELINVO*, no. September, pp. 105–114, 2017.

[17] H. Jain and R. Grover, "Clustering Analysis with Purity Calculation of Text and SQL Data using K-means Clustering Algorithm," *IJAPRR*, vol. IV, no. 44557, pp. 47–58, 2017.

Survey on pedestrian-dynamics models for evacuation process based on game theory

Hocine Benseghir
Department of Engineering Technology
University Malaysia Pahang
Kuantan, Malaysia
Hocine0591@gmail.com

Muhammad Nomani Kabir
Department of Computer Science
University Malaysia Pahang
Kuantan, Malaysia
nomanikabir@ump.edu.my

Mohd Fadzil Mohd Idris*
Department of Engineering Technology
University Malaysia Pahang
Kuantan, Malaysia
*Corresponding author (mohdfadzil@ump.edu.my)

Abdullah Bin Ibrahim
Department of Engineering Technology
University Malaysia Pahang
Kuantan, Malaysia
abi@ump.edu.my

Abstract— **In this paper, we explore the application of game theory to model human decisions making and pedestrian dynamics for evacuation process. In game-theory approach, a pedestrian is represented as a player or an agent where the agent makes its decision by inspecting its neighbors' decision. The game theory is used to study behavioral relations and applied in the science of human behavior decisions. In pedestrian dynamics, the theory deals with individual interaction toward the destination considering psychology and physics. The matrix associated with game theory changes depending on distance to the evacuation exits and obstacles faced on the path, density of evacuees and so on. In this paper, we review the recent articles which studied the game theory to study the pedestrian dynamics and the evacuation process.**

Keywords- Keywords—game theory; pedestrian dynamics; evacuation model; mathematical models; behavior models

I. INTRODUCTION

Pedestrian dynamics is important for crowd management to study the pedestrian movement behavior, which may create congestions resulting in dangerous events e.g., stampede, collision and crush. Specifically, emergency events e.g., fire, explosion and toxic cloud release in a confined area e.g., high-rise building, shopping mall, stadium and industry building can pose extreme danger to the life and property. During emergency evacuation, it is necessary to evacuate the pedestrians from the hazardous area to the safe zone in a shortest possible time [1]. In the event of fire spreading in a building, predicting evacuation path is one of the hardest tasks for the building management [2]. The problem of predicting pedestrian paths and decisions becomes more complex when occupants egress through multiple exits and the groups of evacuees interact in evacuation process. Pedestrian dynamics is used to predict the path, speed and time of evacuees from the start point until the exit point during the evacuation process [3].

Many researchers have developed the evacuation models and software simulators to provide safety of the pedestrians against hazardous events e.g., fire [4]. Thus, there are various evacuation models, approach and software developed for evacuation process, focusing on simulation of the pedestrian flow and the crowd behavior [5]. The software are based on microscopic [6], macroscopic [7] or mesoscopic model [8, 9] to assist the safety managers and building designers to predict the pedestrian movement as well as to study the behavioral refection of the evacuees during their movement.

Game theory was developed by John von Neumann in 1928 to understand situations in which the decision-makers interact [10]. Game theory is a mathematical macroscopic model based on a microscopic model [11], presenting the crowd dynamics in terms of groups, but it is based on a microscopic model that considers pedestrians with rational expectations, where the pedestrians are represented as agents making a decisions [3]. The model can be used to study behavioral relations and evaluate the decision of human behavior. Game theory is used in psychology associated with pedestrian dynamics to deal with individual interaction among pedestrians toward their target, avoiding dangerous area e.g. fire, toxic gas and smoke [10]. In the game theory-based modeling for emergency evacuation, a pedestrian is presented as a player and the player action is executed to reach emergency exits. The model contains psychological parameters to make the evacuation exit choice and decision affected by the individual character [12].

II. GAME THEORY BASED ON SEARCH ALGORITHMS

In this section, we present evacuation process that uses game theory based on the path search algorithms to reduce the egress problem and evacuation route for evacuation plan. Based on decision-making and quantitative risk analysis the first step is to conduct a risk assessment to predict the risk

related to routes movement and path finding. Therefore, game theory deals with the risk scenarios e.g., the interaction of pedestrian and pushing as well crossing to dangerous place.

Figure 1 shows the demonstration of the shortest path-finding method based on the study of interaction of individuals and their decisions. The decisions of the individuals can be evaluated using a game theory model for selecting the appropriate evacuation exit-routes. Search methods e.g., A*search, theta* search and D* search find the shortest path as well as shortest possible time for evacuation.

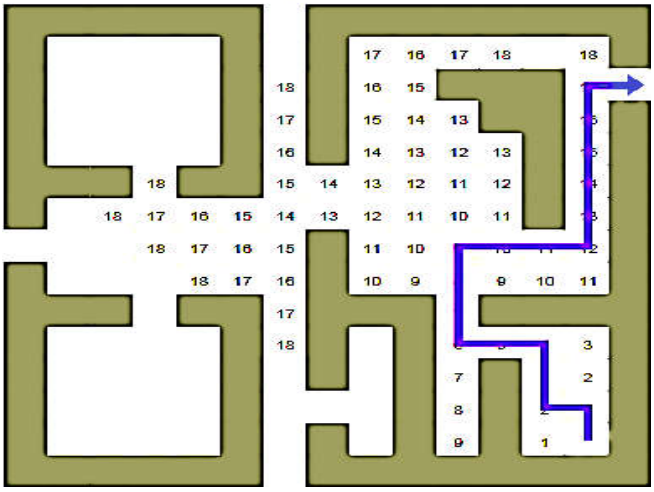


Figure 1. Short path search methods based on decisions and interaction of individuals

A. A* search

During an evacuation process, it is important to carry out fast and safe evacuation by finding the optimal path from hazardous area to a safe exit. Thus, to move from point S(start) to point D (destination), the pedestrians reach the goal by moving around obstacles, walls and blocks. A* search method can be used to search the route choice of the individual pedestrian movement by computing a shortest path. A* uses a cost function to give an estimation of the total path.

B. Theta* search

This method finds the shortest path in a more realistic way by utilizing the angular information. Best path from a start to the goal position is appropriately computed. Researchers developed this method to simulate the pedestrian movement, as well as the robotic application. The method allows dividing the surface with different shapes of grids.

C. D* search

This method is used in the evacuation route planning and mobile robots to find the lowest-cost path through the graph in a real-time scenario. The focused D* algorithm developed by [13] is an efficient search algorithm to deal with large distribution sets [14]. Thus, D* Search is an effective algorithm that can be used to develop route planning for emergency evacuation process.

III. GAME-BASED METHODS

Interactions and behaviors of pedestrians during the emergency situations can be modeled and simulated by a variety of game theory methods. We present a brief overview for the common types of the methods used in this survey.

A. Cooperative and Non- Cooperative

Cooperative and non-cooperative methods are used to study the interaction behavior among the pedestrian conflictions during pedestrians-flow considering evacuation times [15]. This approach is usually used in order to study the group interaction during path finding and to determine the likelihood to change the paths. This model helps understanding if the people follow the signage rules or not. The model can be used in zebra crossing, egress in different places.

B. Symmetric and Asymmetric

Symmetric game type has more acceptability where the player has an equal chance for deciding to travel. If the player's identity can be changed without changing the payoff to the strategies, then the game is symmetric. On the contrary, the game is asymmetric where there are no identical strategy sets for both players.

C. Zero-sum and Non-zero-sum

In zero-sum model, the sum of the payoffs of the two players is zero for any choice of strategies which means one player wins exactly in the same extent that the other player loses [16]. Meanwhile, the other type such as traveler's dilemma is non-zero-sum games. The model is used to study the route choice problem as a zero-sum game to find the lane-changing for connected vehicles [17], The model can be used for path choice and exit egress selection.

D. Mean field games

This type for game theory is used to study strategic decision making for population interactions of pedestrian crowd. Some applications of mean-field-type include route traffic networks, multi-story building evacuation, power distributions networks and so on[11].

IV. GAME THEORY IN PEDESTRIAN DYNAMICS

Pedestrian dynamic models in the context of the evacuation process such as game theory, cellular automata and social force model are used for comprehensive understanding of the pedestrian movements. Pedestrian dynamics models divided into two main types (i) microscopic models e.g., social force model [18, 19], cellular automata [20-24] which study people movement and can include interaction between people as each pedestrian occupies one circle or cell which may be partially occupied by the neighbors. (ii) macroscopic models e.g., continuum

model, queueing model [25], game theory models [10] and fluid dynamic models [26]. However, our scope is to study the models that utilize the game theories that deal with strategic and social interactions among the pedestrians during an evacuation process.

Figure 2 presented the steps of the generalized evacuation method using a game theory - by starting from the initial response until the destination by taking time into account.

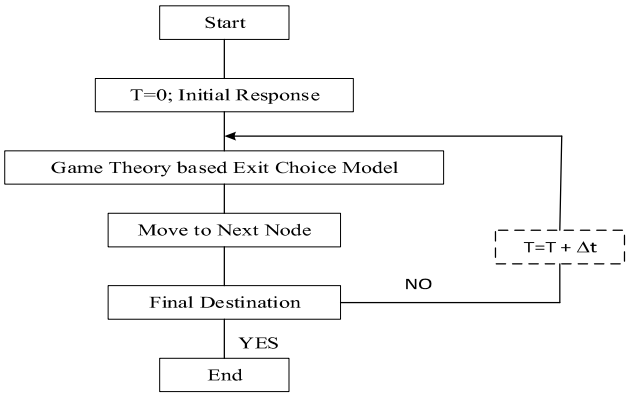


Figure 2. Evacuation process for each person[10]

Guan et al. [27] integrated a cellular automata approach with a snowdrift game theory for pedestrian behavior to simulate conflictions associated with the crowd dynamic during emergency evacuation process. The study consists of two types of individuals - cooperators and non-cooperators and with two main factors - cost coefficient and fear index. The test results reveal that the cost coefficient and fear index increased the evacuation time. Moreover, the ratio of non-cooperators to cooperators exhibits steady states even with different values of parameters, due to the effects of self-organization. Tanimoto et al. [22] studied the effects of bottleneck and evacuation exit during emergency by integrating a cellular automaton model for pedestrian dynamics with game theory and field approximation analysis. In this study, the authors compared between the pedestrian problems faced in a bottleneck exit using the collision effect derived from game theory and static floor field. The bottleneck problem, in that pedestrians rush toward the exits, results in a breakdown of pedestrian flow from the exit due to congestion effect. The authors focused on bottleneck problems in pedestrian flow as they are theoretically relevant to game theory. The test results demonstrated that pedestrian flow rate through an exit is enhanced by setting a proper obstacle before the exit to reduce collision probability at the exit.

Lo et al. [10] integrated exit choice model with game theory for evacuation during escape from a fire zone in building using the dynamic exit choice process of evacuees. The authors studied the evacuees' interaction and their effects for evacuation path and rescue time for a multi-exit area. The author used non-cooperative type of game theory to deal with individual interactions among pedestrians toward their goals (exit). Results presented that the approach will guide the evacuees interaction with the environment, by scrutinizing the familiarity effects and environmental stimuli on exit selection process. Bjørnskau [28] used a game theory to analyze the behavior interaction among pedestrians, car drivers and cyclists at crosswalk in Norway during normal situation of traffic. They studied the crossing behavior at three types of crosswalk in order to validate the study. The results showed that the game theory model can be a valuable tool to understand and predict interactions of road user in order to improve traffic safety. Mesmer et al. [29] incorporated an individual's decision making and game theory approach with a velocity obstacle method to capture human behavior during maneuvering in an environment occupied by obstacles or other individuals. The model can be used to examine decision-making process for the human dynamic captures, instead of modeling the behavior directly. The model was used in different case-studies and examples to show that the decision-based method is effective to simulate the human behaviors during evacuations and in crowded situation. Talebpour et al. [17] used a lane-change model based on game theory for connected vehicles to study the drivers' behavior during driving and to enhance drivers' operational decisions e.g., acceleration, lane-changing and awareness of the surrounding traffic. Results showed that the proposed model can be used to determine the lane-changing behavior with some limitations. von Schantz et al. [30] developed a computationally light evacuation simulation model for pedestrian movement. The authors incorporated cellular automaton with game theory in order to enable the agents with decision-making capacities. In the game theory, an individual observes the decision of its neighbors and reacts according to its movement. In the game matrix, the agents have decisions to choose: (i) patient or (ii) impatient. Moreover, the expected rescue time and available safe egress time are taken as decision of the agents. The results presented that impatient agents can reach the evacuation exit faster than patient agents. The results also showed that more agents were impatient and the evacuation was slower for the whole crowd. Wang et al. [31] presented a game-theory approach for predictive lane-changing for automated driving systems. The parameters taken for the presented model are based on the game theory where the controlled vehicles decisions change depends on the other vehicles decision. The results showed that the proposed approach is useful for an effective lane-changing movement. Li et al. [32] integrated game theory approach with Monte-Carlo method for optimization evacuation routing in complex scenarios. The evacuation model considers two steps, where - the first step uses a Monte Carlo method to assess the risk level of the route, the second step uses a novel model based on game theory to compute a solution that guides the evacuation of all occupants from the selected area. According to the authors, the model presented in the study can guide evacuees to the route. Mesmer et al. [33] developed a game theory model to determine egress decision and exit choice during an emergency evacuation. The developed model can be used to study exit decision making that captures behaviors and interactions among the pedestrians. The model was validated by comparing to the experimental data obtained from a real case-study. According to the authors, the proposed model

provides more reliability than the traditional large-scale heuristic methods. Bouzat et al. [34] used a game-theory-based cooperative and non-cooperative model with a lattice gas model to simulate the pedestrian evacuation. The model was used to study the evacuation process of pedestrians from a room with a single exit. The developed model is used to determine individual decisions that may influence to the interactions of pedestrian movement. The results showed that, due to mutual cooperation, the evacuation process of cooperators is accelerated. Thus, evacuation with cooperators becomes than that with non-cooperators. Shi et al. [35] used a snowdrift game theory to simulate interaction among the conflicting pedestrians during evacuation from a room with a single exit. The authors integrated the game model with a modified lattice gas model to study pedestrian movement. The payoff matrix associated with the game theory represents the interaction among the pedestrians along their way to exit. The test results demonstrate that each person interacting with its neighbors randomly endeavors to take over the empty neighbors' position. Moreover, the integrated model exhibits different dynamics before leading to a consistent state. Test results shows that the escape time becomes shorter, when the density of pedestrians is lower Chizari et al. [36] integrated an agent-based approach with game theory model for modeling evacuee uncertainty behavior during evacuation in building. The results of the study showed that with different evacuation scenarios, time of evacuation change when they take uncertainty into consideration. Braglia et al. [12] Presented a game theory for the evacuees in egress exit during emergency situation. The presented model consists of aspects and psychological parameters to obtain a suitable presentation of the human behavior and evacuation management during emergency situations. Zheng et al. [15] Integrated a game theory with an evacuation model to determine the density of evacuees and their evacuation times to reach exit in a big room size with a single exit door. The model is based on evacuation dynamics where the cellular automata used to determine evacuee's movement rules and the game-theoretical model used to determine the evacuee's behaviors (cooperative and competitive) rules. The results showed that (i) cooperation among evacuees' decreases when the urgency of evacuation increases; (ii) in the evacuation, the hyper-rationality between evacuees reduces evacuation efficacy and delays evacuation times; (iii) the imitation effect increases cooperation between evacuees, meanwhile, it decreases evacuation efficacy and increases evacuation times. Lachapelle et al. [11] presented a pedestrian crowd approach based on game theory approach to model crowd motion and congestion behavior. The developed approach is used to describe anticipation phenomena and to reduce computational cost of macroscopic simulations. The model presented a mean field game where pedestrians are players with decision and they are willing to improve their paths and reduce congestion with certain costs. The model uses Mean Field Games (MFG) and focuses on interactions between two groups of pedestrians e.g., uncooperative and congestion behavior. The authors presented numerical solvers including complex group behaviors and anticipation phenomena e.g., symmetry breaking, lane formation to reproduce characteristic features of crowds. The model was validated by presenting different examples. According to the authors, the proposed models give a good effect of the possibilities of mean field game theory.

The outcome of our paper is to review all the previous studies that implement game theories for pedestrian dynamics for evacuation process. Table 1 presents a comparison of the applications of the game theory for pedestrian dynamic during normal and emergency evacuation situations studied on the above reviewed works. The second column shows the different types of game theory models integrated with cellular automaton or agent-based models. The third column provides the scenario of the study. The fourth column lists the players involving in the decision-making of the game theory, where interaction of the players is used to study the behaviors and movements of the individuals. The last column shows the results or outcomes from the study.

TABLE I. COMPARISON OF REVIEWED WORKS ON THE APPLICATIONS OF THE GAME THEORY FOR PEDESTRIAN DYNAMICS

Author	Type	Scenario	Player	Outcomes
Guan et al. [27]	- Cellular automaton -cooperation/defection game theory (GT)	During evacuation processes	Pedestrians	Simulate crowd dynamic conflicts
Tanimoto et al. [22]	- Cellular automaton -cooperation/defection GT	Bottleneck effect at an evacuation exit during emergency	Pedestrian	Predict evacuation time to safe exit
Lo et al. [10]	Cooperation/defection GT	During escape from a fire zone in building (Exiting pattern for a 2 room with 2 exit situations)	Pedestrian	Studied the evacuees' interaction and their effects for evacuation pattern and clearance time
Bjørnskau [28]	Leader game	At zebra crossings	Cyclists / car drivers	Analyzed the interaction
Mesmer et al. [29]	Vacate-GT	In an environment occupied by obstacles or individuals	Human	Capture interactions during maneuvering
Talebpour et al. [17]	Two-person non-zero-sum non-cooperative	Lane-changing behavior	Connected vehicles	Predict lane change
von Schantz et al. [30]	Cooperative / Non-cooperative GT	Integrated game theory with a cellular automaton	Pedestrians	Simulate behavior of the crowd during evacuation
Wang et al. [31]	Cooperative	Predictive lane-changing using game theoretic	Vehicles	Predict lane change
Li et al. [32]	Game theory Based on Monte Carlo	Congestion problem in complex structures	Pedestrians	Optimize evacuation route
Mesmer et al. [33]	-Bayesian/ Pseudo-Bayesian games -Vacate GT	Human behaviors and decisions in egress	Evacuees	Study exit decision making
Bouzat et al. [34]	Cooperation/defection GT	Room with a single door	Pedestrians	Analyzed the interaction of pedestrian during evacuation
Shi et al. [35]	A snowdrift GT	A single room with one door during evacuation	Pedestrians	Simulate interaction of pedestrian
Chizari et al. [36]	Agent-based approach with GT	Uncertainty pedestrian behavior during evacuation in building	Occupants	Modeling evacuees
Braglia et al. [12]	Game theory approach	Evacuees exit selection in emergency conditions	Evacuees	Study human behavior during emergency situations
Zheng et al. [15]	-Cooperative / Non-cooperative GT -cellular automata (CA	Exit from room with an exit door	Evacuees	Predict evacuation time and make decision
Lachapelle et al. [11]	Mean field GT	crowd situation, uncooperative and congestion behavior	Groups of pedestrians	Model crowd motion and congestion behavior
Bjørnskau [37]	Game theory	Pedestrian crossings as a co-ordination game	Cyclists /vehicles	Predict the behavior and decision of cyclists and vehicles
Tay [38]	Game theory	Accident involving small and large cars	Vehicles/ Drivers	Study risk of driver fatality related to driver's decision for choosing type of the cars

V. DISCUSSION AND CONCLUSION

This paper presented a survey on game theory models that are used for pedestrian dynamics. The paper covers the most recent articles published in the area of pedestrian movements to their goal (evacuation exit) during normal and emergency situations. The articles which present the new models using a game theory to study the people movement are discussed. Game-theory models can be utilized to study complicated interactions among people and their behavior. Game-theory models are developed to explain people's choices, decision-making and the informal norms of behavior. The integrated models such as cellular automata with cooperative and non-cooperative game theory are used to study the cooperative and non-cooperative people during the evacuation, providing powerful techniques for analyzing complex pedestrian dynamics and transport systems. We provide a comparison of features, scenario and game theory matrix of the most interesting articles. The game models which were presented in the article showed that the players have full knowledge of their own decisions and payoffs, as well as knowledge of the decisions and payoffs of the other players. This survey is presented in order to make brief and precise review in the field of pedestrian dynamics for evacuation process.

ACKNOWLEDGMENT

This research is supported by a project no. RDU1703118 from Universiti Malaysia Pahang

REFERENCES

[1] M. N. A. Khalid and U. K. Yusof, "Dynamic crowd evacuation approach for the emergency route planning problem: Application to case studies," *Safety Science,* vol. 102, no. Supplement C, pp. 263-274, 2018/02/01/ 2018.

[2] N. E. Groner, "A decision model for recommending which building occupants should move where during fire emergencies," *Fire Safety Journal,* vol. 80, pp. 20-29, 2// 2016.

[3] H. Vermuyten, J. Beliën, L. De Boeck, G. Reniers, and T. Wauters, "A review of optimisation models for pedestrian evacuation and design problems," *Safety Science,* vol. 87, no. Supplement C, pp. 167-178, 2016/08/01/ 2016.

[4] Z. Zainuddin and M. Shuaib, "Modification of the Decision-Making Capability in the Social Force Model for the Evacuation Process," *Transport Theory and Statistical Physics,* vol. 39, no. 1, pp. 47-70, 2010/01/13 2010.

[5] X. Zheng, T. Zhong, and M. Liu, "Modeling crowd evacuation of a building based on seven methodological approaches," *Building and Environment,* vol. 44, no. 3, pp. 437-445, 2009/03/01/ 2009.

[6] D. R. Parisi and C. O. Dorso, "Microscopic dynamics of pedestrian evacuation," *Physica A: Statistical Mechanics and its Applications,* vol. 354, pp. 606-618, 2005/08/15/ 2005.

[7] I. A. Ndiaye, E. Neron, and A. Jouglet, "Macroscopic evacuation plans for natural disasters," *OR Spectrum,* journal article vol. 39, no. 1, pp. 231-272, 2017.

[8] M. Shi, E. W. M. Lee, and Y. Ma, "A novel grid-based mesoscopic model for evacuation dynamics," *Physica A: Statistical Mechanics and its Applications,* vol. 497, pp. 198-210, 2018/05/01/ 2018.

[9] M. Shi, E. W. Ming Lee, and Y. Ma, "A Newly developed Mesoscopic Model on Simulating Pedestrian Flow," *Procedia Engineering,* vol. 211, pp. 614-620, 2018/01/01/ 2018.

[10] S. M. Lo, H. C. Huang, P. Wang, and K. K. Yuen, "A game theory based exit selection model for evacuation," *Fire Safety Journal,* vol. 41, no. 5, pp. 364-369, 2006/07/01/ 2006.

[11] A. Lachapelle and M.-T. Wolfram, "On a mean field game approach modeling congestion and aversion in pedestrian crowds," *Transportation Research Part B: Methodological,* vol. 45, no. 10, pp. 1572-1589, 2011/12/01/ 2011.

[12] M. Braglia, D. CASTELLANO, and R. GABBRIELLI, "A NOVEL GAME THEORY BASED EXIT SELECTION MODEL IN EMERGENCY CONDITIONS," *Advances in Complex Systems,* vol. 16, no. 07, p. 1350018, 2013.

[13] A. Stentz, "The focussed D^* algorithm for real-time replanning," in *IJCAI*, 1995, vol. 95, pp. 1652-1659.

[14] Y. Matsubara, Y. Sakurai, and M. Yoshikawa, "D-Search: an efficient and exact search algorithm for large distribution sets," *Knowledge and Information Systems,* journal article vol. 29, no. 1, pp. 131-157, October 01 2011.

[15] X. Zheng and Y. Cheng, "Modeling cooperative and competitive behaviors in emergency evacuation: A game-theoretical approach," *Computers & Mathematics with Applications,* vol. 62, no. 12, pp. 4627-4634, 2011/12/01/ 2011.

[16] N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, *Algorithmic Game Theory*. Cambridge University Press, 2007.

[17] A. Talebpour, H. S. Mahmassani, and S. H. Hamdar, "Modeling lane-changing behavior in a connected environment: A game theory approach," *Transportation Research Part C: Emerging Technologies,* vol. 59, pp. 216-232, 2015/10/01/ 2015.

[18] D. Helbing and P. Molnár, "Social force model for pedestrian dynamics," *Physical Review E,* vol. 51, no. 5, pp. 4282-4286, 05/01/ 1995.

[19] D. Helbing, I. Farkas, and T. Vicsek, "Simulating dynamical features of escape panic," *Nature,* vol. 407, p. 487, 09/28/online 2000.

[20] S. Wei-Guo, Y. Yan-Fei, W. Bing-Hong, and F. Wei-Cheng, "Evacuation behaviors at exit in CA model with force essentials: A comparison with social force model," *Physica A: Statistical Mechanics and its Applications,* vol. 371, no. 2, pp. 658-666, 2006/11/15/ 2006.

[21] I. Santé, A. M. García, D. Miranda, and R. Crecente, "Cellular automata models for the simulation of real-world urban processes: A review and analysis," *Landscape and Urban Planning,* vol. 96, no. 2, pp. 108-122, 2010/05/30/ 2010.

[22] J. Tanimoto, A. Hagishima, and Y. Tanaka, "Study of bottleneck effect at an emergency evacuation exit using cellular automata model, mean field approximation analysis, and game theory," *Physica A: Statistical Mechanics and its Applications,* vol. 389, no. 24, pp. 5611-5618, 2010/12/15/ 2010.

[23] R. Alizadeh, "A dynamic cellular automaton model for evacuation process with obstacles," *Safety Science,* vol. 49, no. 2, pp. 315-323, 2011/02/01/ 2011.

[24] X. Zheng and Y. Cheng, "Conflict game in evacuation process: A study combining Cellular Automata model," *Physica A: Statistical Mechanics and its Applications,* vol. 390, no. 6, pp. 1042-1050, 2011/03/15/ 2011.

[25] S. P. Hoogendoorn, F. van Wageningen-Kessels, W. Daamen, D. C. Duives, and M. Sarvi, "Continuum Theory for Pedestrian Traffic Flow: Local Route Choice Modelling and its Implications," *Transportation Research Procedia,* vol. 7, pp. 381-397, 2015/01/01/ 2015.

[26] I. Mohd, A, I. Venkat, and P. D. Wilde, "Uncertainty in a spatial evacuation model," (in English), *Physica A: Statistical Mechanics and its Applications,* Article vol. 479, pp. 485-497, 2017.

[27] J. Guan, K. Wang, and F. Chen, "A cellular automaton model for evacuation flow using game theory," *Physica A: Statistical Mechanics and its Applications,* vol. 461, no. Supplement C, pp. 655-661, 2016/11/01/ 2016.

[28] T. Bjørnskau, "The Zebra Crossing Game – Using game theory to explain a discrepancy between road user behaviour and traffic rules," *Safety Science,* vol. 92, pp. 298-301, 2017/02/01/ 2017.

[29] B. L. Mesmer and C. L. Bloebaum, "Modeling decision and game theory based pedestrian velocity vector decisions with interacting individuals," *Safety Science,* vol. 87, pp. 116-130, 2016/08/01/ 2016.

[30] A. von Schantz and H. Ehtamo, "Spatial game in cellular automaton evacuation model," *Physical Review E,* vol. 92, no. 5, p. 052805, 11/09/ 2015.

[31] M. Wang, S. P. Hoogendoorn, W. Daamen, B. van Arem, and R. Happee, "Game theoretic approach for predictive lane-changing and car-following control," *Transportation Research Part C: Emerging Technologies,* vol. 58, pp. 73-92, 2015/09/01/ 2015.

[32] W. Li, J. Zhu, H. Li, Q. Wu, and L. Zhang, "A game theory based on Monte Carlo analysis for optimizing evacuation routing in complex scenes," *Mathematical Problems in Engineering,* vol. 2015, 2015.

[33] B. L. Mesmer and C. L. Bloebaum, "Incorporation of decision, game, and Bayesian game theory in an emergency evacuation exit decision model," *Fire Safety Journal,* vol. 67, pp. 121-134, 2014/07/01/ 2014.

[34] S. Bouzat and M. N. Kuperman, "Game theory in models of pedestrian room evacuation," *Physical Review E,* vol. 89, no. 3, p. 032806, 03/14/ 2014.

[35] D. Shi and B.-H. Wang, "Evacuation of pedestrians from a single room by using snowdrift game theories," *Physical Review E,* vol. 87, no. 2, p. 022802, 02/04/ 2013.

[36] H. Chizari *et al.*, "Agent-based approach for modeling evacuee uncertainty behavior using game theory model," *Life Science Journal,* vol. 10, no. 3, pp. 1350-1355, 2013.

[37] T. Bjørnskau, "Spillet om gangfeltet - 6 år etter," *Samferdsel,* vol. 4, pp. 22-23, 2007.

[38] R. Tay, "The prisoner's dilemma and vehicle safety: Some policy implications," *Journal of Transport Economics and Policy,* pp. 491-495, 2002.

Action Training and Assessment System with Kinect

Ivana Valentine Masala
Informatics Engineering Department
De La Salle Catholic University Manado, Indonesia.
Kombos Kairagi I 95000
imasala@unikadelasalle.ac.id

Shih-Lin Wu
Dept. of Computer Science and Information Engineering
Chang Gung University, Taiwan.
Tao-Yuan 333
slwu@mail.cgu.edu.tw

***Abstract*—To capture the action of the user in the action training and assessment system that will be developed, this work use Kinect as a means of data retrieval. The system is set to capture simple action and complicated action for comparison. The first action is a parameter used in comparison with the user, where the recording is taken from the professional user who has been stored in the database system. Furthermore, the second action is an action that comes from a novice user who performs the action in the current time and the result will be compared with the first action. The action data takes from X – Y coordinates value. The obtained the value of X and Y from the shoulder to the palm. The system proposes with our algorithm to show the assessment result of both actions comparison. It is displayed in the compatible percentages and the opposite one.**

Keywords—Actions, Kinect, Data Retrieval, Assessment.

I. INTRODUCTION

Nowadays, Kinect is used as a learning tool for children. As an efficient and attractive educational tool, which is not only a virtual but in real time for the development of children's motor in system learning, even the children can practice the things learned. Previously, Kinect development in the area of healthcare promotion and exercise training system development are very crowded. Furthermore, the changes were made by Kinect developers by evaluated the improvement of technology and get the new ideas to adopt the smaller Kinect concept that can be adjusted on the computer directly.

Our motivation in this research is to build a system that can be used as a training system for the action of movement which is done by the command provided by the system. The existing system combined with Kinect. The action is a simple action and complicated action of dance. With the assessment made by the system, it can give the comparison results of two actions of movement in the form of the percentages.

The virtual tutor is one of the nice alternatives combine system. Some applications developed with virtual tutor concept and in real life, it uses as a home-based system. A home-based system uses to learn some exercise or rehabilitation training. Currently, the dance exercise and health promotion for rehabilitation training are using Kinect.

This work address to design an action training and assessment system with Kinect. First of all, it need to retrieve the data from a professional user, in the case as a parameter motion. This data retrieval will be stored in the system. Further, a real-time user will perform the action by following the data that shown the track on the screen, and record the obtained data. Afterward, both data will compare and show the compatibility result between the two actions.

Simple action and complicated action is a part of action dance. The virtual tutor built by Unity3D uses the data retrieval. Besides, the user real-time screen, show through the OpenNI and OpenCV platforms. Considering some parameters in this work, the author decide to take those from the time period of the action, coordinate points, the action path of each coordinates distance, and the action error.

Contribution in this works which is, the data retrieval and user real-time data captured by Kinect, generally from the skeleton algorithm from Kinect show the joints of the human body in the screen. The virtual tutor is a 3D model generated from Unity3D using first data retrieval action. As a training and assessment of two actions, user real-time data adjust to get the compatible percentages in addition the user real-time will perform the right action. In this case, a part of the action dance needs to perform in the correct way. The measurement to assess until we get the percentages result based on the parameters.

The paper consist of some section such as introduction in the first section, some related works in the second section, the system framework and the method in the third section, fourth section showing the result and the fifth section is conclusions.

II. RELATED WORKS

A. Training System and Kinect Device

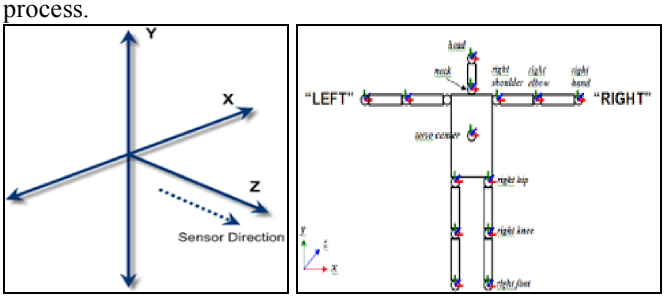
Dance exercise is one of the exercise training systems. Joana and Filipa [1] had the dance exercise explanation in the health promotion field of healthy aging. Dance movement was possible to change someone's mood into a relaxing stage and decreased the aging factor by performed the dance movement. Dance as physical exercise form decrease anxiety explained by Aniko et al. [2] it increased self-esteem and improved psychological well-being. Dance exercise with virtual reality effect in health promotion field, through the patients of Parkinson disease of the daily activities, explained by Nam-Yong et al. [3]. The Parkinson's disease is an unstable balancing of someone's daily activities. The positive effects of virtual reality dance exercise on the patients, balancing their body activities. Chi-Jui et al. [5] had the method to track the human joints skeleton position with depth object sensing detection of multiple Kinects.

B. *Evaluation, Assessment, and Motion Capturing System*

Min-Chun et al. [4] was a system that divides to 2 parts of dance performances. Retrieval data captured by Kinect and assessed by the system. Virtual and real-time environments measured the differences of each joint in human poses. Dynamic Time Warping (DTW) method modified by Slim et al. [6] in another way, the author have evaluated the performances of some actions. Their proposed a dance performance online system and provided the interactions between two users as a dance teacher and as a student. As previously, here Li and Shao [7] explained the gesture recognition algorithm. Based on the DTW (Dynamic Time Warping) and bit modified matching experiment between two actions, their applied in action recognition training. Another proposed dance motion synthesize by following the music content, Fan et al. [8]. Music background used to find the relation of the coefficient correlation. Dynamic programming procedure generated the sequence of dance motion segments. To apply the virtual reality concept on the educational field as a purpose from Mingshao Zhang et al. [9] through Kinect game-based they developed with three aspects to capture the motions such as measuring tool, acquiring range data and tracked the object, also recognized the object's voice. To improve the motions capturing, Bjorn Muller et al. [10] provided the VICON system as a step motion capturing system was applied with their method. Flexibility and high accuracy on some parameters comparable, shown the VICON system got the result motion collaborated by Kinect v2.

III. System Framework and Proposed Method

Virtual tutor data retrieve from the professional user and store the data to the system, in another side it adjust to the 3D model to build the 3D virtual tutor. User real time is a novice user who will perform the action by following the virtual tutor in the system. This section is showing the tools in our architecture system and the flow of the system model process.



(a) (b)

Fig. 1. (a) 3D measurement plane by Kinect sensor, (b) 15 joints the human skeleton.

The position of X Y Z to get the real world value is explained by the figure 1. Positive X value defined the right side of the device, positive Y value in the upside and Z value measures the depth of the distance between Kinect and the object with respect to the Kinect device. 15 joints of the human body skeleton data position generated by 3D plane X Y Z, the data position measured in millimeter (mm).

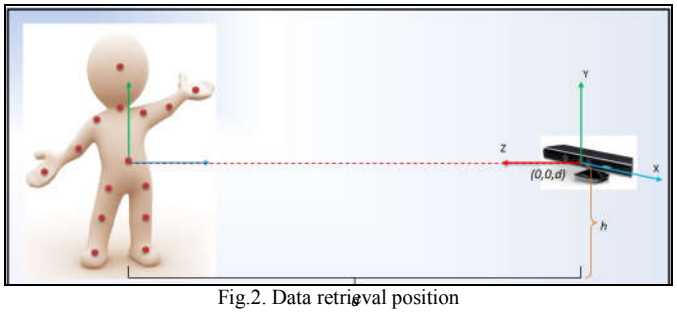


Fig.2. Data retrieval position

In order to retrieve the data action of the professional person as a real coach called a virtual tutor as shown in the figure. User data action for the current time was taken in the same way. The data obtained through this ground position. It manually set between the device distance and object, where the distance d=2,5m and the height of the Kinect device from the ground h=1m. Considering the center of the human body need to adjust to the same line with the Z plane of the Kinect.

Afterward, to obtain the data action of the users, it need to store the coordinate's X – Y value of the palm joint that respects to the shoulder. This action based on the user's action captured by the palm joint. Here, the system will calculate the ground of the coordinates that is provided screen of the system as in figure 3.

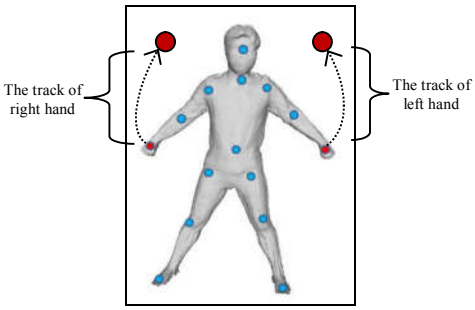
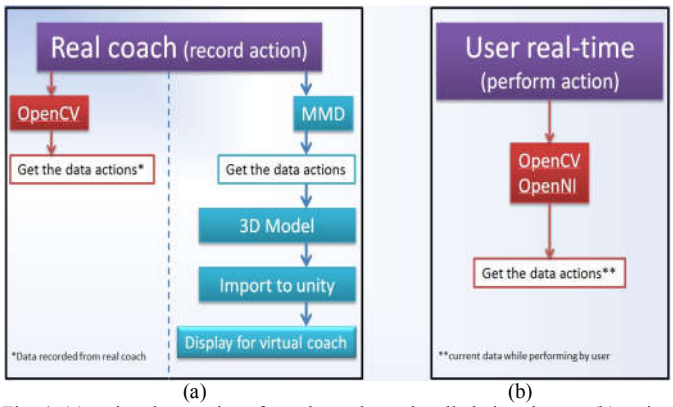


Fig.3. obtained the data coordinate action in the X – Y plane from palm joint.

To detect the shape of user, still based on the skeleton algorithm to show the center of the coordinates. The data of X – Y coordinates obtained as the palm coordinates lines. Its respect to the user's shoulder position as the center. The measurement of the user's shape, in this work used fix shape as a professional user.



(a) (b)

Fig. 4. (a) action data retrieve from the real-coach called virtual tutor, (b) action data of the real-time user.

B. User's Category

The virtual tutor is a 3D model that running and perform the retrieval data from a professional person. The dance actions were recorded and put into the 3D model through a software --MMD (Miku-Miku Dance)-- which play in the role for recorded data. The dance action of the virtual tutor as an instruction to the real-time user.

Real-time user as the student who is novice about the action. By performs the action and follow the virtual tutor in the system, the real-time user generates the current data action. The real-time user can see the position of their action. It starts from the initial position by following the instruction. Adjustment of the both hands before starts the action.

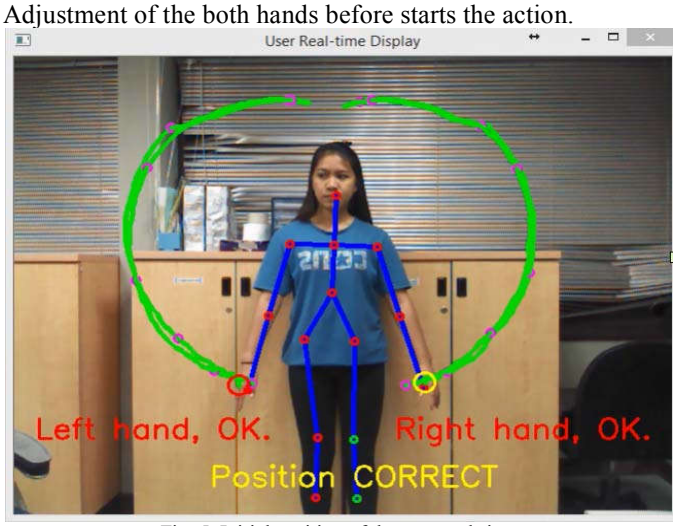


Fig. 5. Initial position of the user real-time

C. System Architecture

The system architecture likely showed in the figure 5. With each function and process used the method to applied in the system and finished with completed some demos.

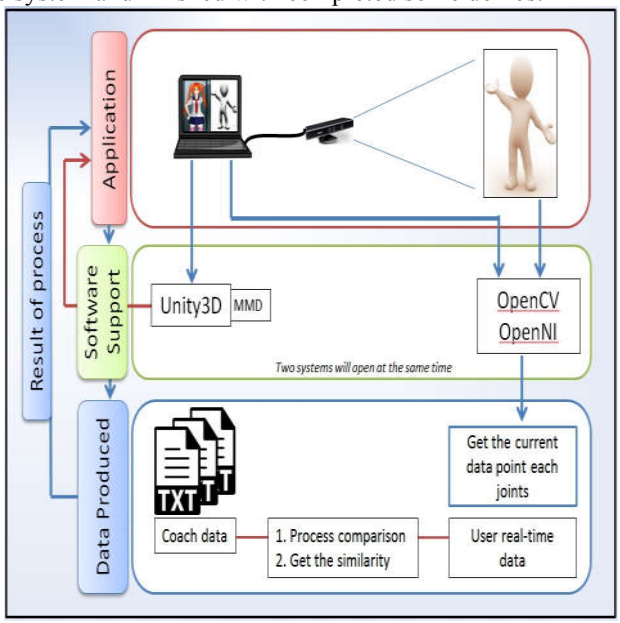


Fig. 6. System Architecture

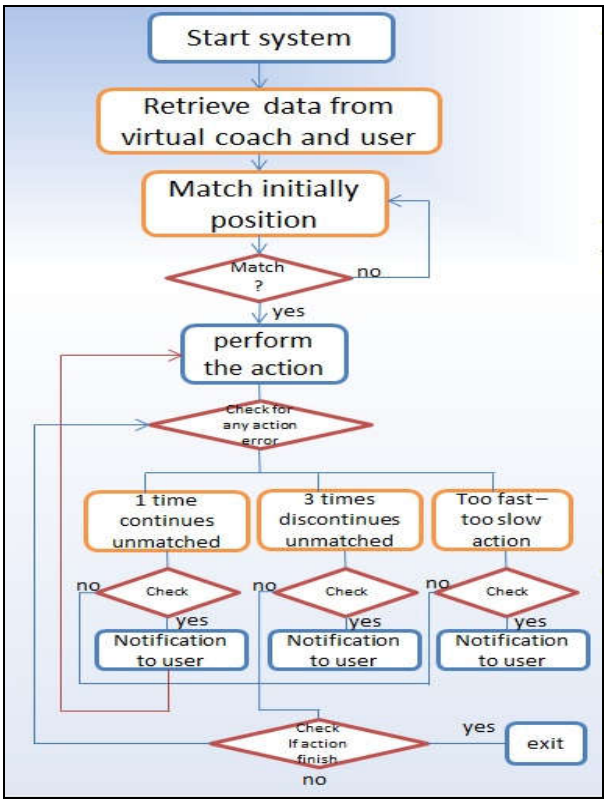


Fig. 7. System flow by applying the algorithm

Our proposed method was the way to find the similarity between the two actions. The comparison algorithm processed the parameters considered the time, coordinates, path length, and the rest of the flow error of actions in the system. Afterward, explained the way to get compatible percentages of the actions and the show the scoring result value.

D. The similarity of Two Actions

To measure the similarity of two actions, many ways are there through a measuring method such as general similarity function that used sequence alignment, vector and signal encoding, time warping, etc. Previously [4] presented a similarity measurement between two poses with modified DTW (Dynamic Time Programming) method. Here, we present our method to find the similarity of that different data action with comparison algorithm in our system. Furthermore, the author used 4 parameters such as; time, points, path length, and flow error.

E. Matching Actions and Scoring Result

This work used a number of 30 as a value of the distance ratio to define the threshold (based on 30fps). Scoring Percentage calculated by the average value, and the total match and total unmatched with the conditions of percentages of matched points = 1, else = 0.

$$compatible\ action = \frac{\sum_{i=1}^{n} compatible\ data}{all\ data} * 100\%$$

$$incapable\ action = \frac{\sum_{i=1}^{n} incapable\ data}{all\ data} * 100\%$$

Table 1. Scorinng

Comparing the Data	t					Result	Score (%)
	t_1	t_2	t_3	t_4	t_5		
Viirtual tutor data*(x)* and user real time (x')	1	0	1	1	1	matched	86.3%
	1	0	0	1	1	matched	61.8%
	0	0	0	0	0	unmatched	95.4%

IV. Experiment and Result

The figure shows the experiment through the way of the action recorded. The real-time user performs the action by following virtual coach action. The simple action is Slow Tap Dance that focused on timing and beats [11]. Furthermore, compare both data and calculated based on the parameters. Some figure showed the experiment result.

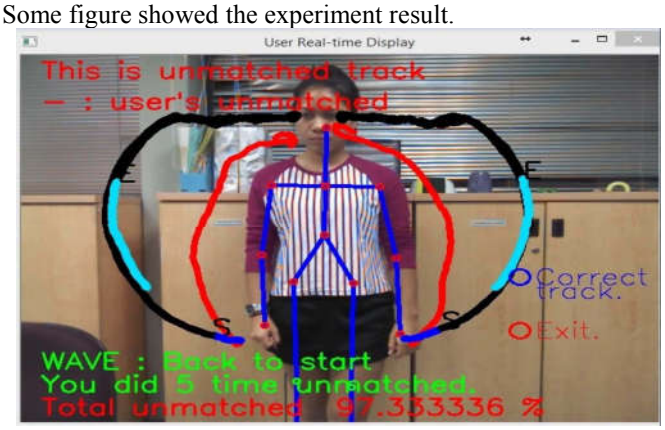


Fig.8. Option's of the error action

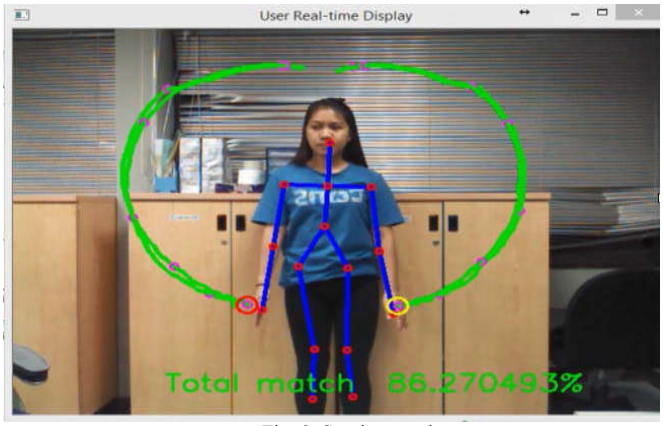


Fig. 9. Scoring result

Figure 7 shown the user real time did the wrong action is in the red line, and the system guide according to the steps. If user choose to correct the action, by the hand user can swap the selection option as show in the figure. Besides, figure 8 shown the result when the user perform the right action without any error.

V. Conclusion

This system compares both data generated by the palm joint of the human body. By following the track of the action, the system can train a novice user through the instruction that is shown on the screen of the system. With Kinect home-based system as tools combine with some soft wares, show the flow error to the real-time user who can make the correct action. The system processes the action data and shows the compatible scoring. The efficiency of low-cost tool helped the user to improve dance exercise at the home by themselves. This work apply in dance exercise simulation, as an example in simple action that is a part of the dance action.

The variety of the action also can apply in the system such as badminton training action for a beginner, Taichi action, and selected a rehabilitation health care action, etc. For the future work, we will improve the system to minimize the deficiency of the system, improve the calculation of the similarity between more than two actions, and add the ration calculation to detect more dinamic user's body shape. At the last, this work will improve with another tool which is Intel Real-sense with another technology.

References

[1] B. R. Joana and F. N. Filipa, "Dance Movement Therapy as a Health Promotion Tool in Aging", Universitat Autonoma de Barcelona, February 2016.

[2] M. Aniko, K. Orsolya, U. Robert, G. Mark D, D. Zsolt, "Why Do You Dance? Development of the Dance Motivation Inventory (DMI)", PloS one 2015 – journals.plos.org, March 2015.

[3] L. Nam-Yong, L. Dong-Kyu, S. Hyun-Seung, "Effect of Virtual Reality Dance Exercise on the Balance, Activities of Daily Living, and Depressive Disorder Status of Parkinson's Disease Patients", Journal of Physical Therapy Science, pg. 145-147, January 2015.

[4] H. Min-Chun, C. Chi-Wen, C. Wen-Huang, C. Che-Han, L. Jui-Hsin, and W. Ja-Ling, "Real-Time Human Movement Retrieval and Assessment With Kinect Sensor", IEEE Transaction On Cybernetics, Vol. 45, No. 4, April 2015.

[5] W. Chi-Jui, Z. Zhou, C. Yizhe, A. El-Sayed, E. Sven, and C. Constantin, "Out of Sight: a toolkit for tracking occlude human joint positions", Springerlink.com, December 2016.

[6] E. Slim, A. Dimitrios, T. Robin, G. Marc, K. Philip, M. David, D. Petros, D. Angelique, and O. Noel. E., "An Advance Virtual Dance Performance Evaluator", 2012 IEEE International Conference of Acoustics, Speech and Signal Processing (ICASSP), pg. 2269-2272, Kyoto, March 2012.

[7] L. Nianfeng, S. Yanhiu, "Study on Action Recognition Based on Kinect and Its Application in Rehabilitation Training", 2015 IEEE Fifth International Conference on Big Data and Cloud Computing (BD Cloud), pg. 265-269, Dalian, August 2015.

[8] F. Rukun, X. Songhua, G. Weidong, "Example-Based Automatic Music-Driven Conventional Dance Motion Synthesis", IEEE Transaction on Visualization and Computer Graphics, Vol. 18, pg. 501-515, March 2012.

[9] Z. Ming Shao, Z. Zhou, C. Yizhe, A. El-Sayed, E. Sven, and C.Constantin, "Recent Developments in Game-Based Virtual Reality Educational Laboratories Using the Microsoft Kinect", iJet-Vol. 13, No. 1, 2018.

[10] M.Bjorn, I. Winfried, G. Martin A. L. Nicolas, "Improved Kinect Sensor Based Motion Capturing System for Gait Assessment", dx.doi.org, January 2017.

[11] "Types of Dance – Categories", http://www.dancefacts.net/dance-types/types-of-dances/, May 2015.

Customer Behavior Analysis Using Data Mining Techniques

Patcharin Ponyiam
Department of Computer Science
Khon Kaen University
Khon Kaen, Thailand
e-mail: pacharin_po@kkumail.com

Somjit Arch-int
Department of Computer Science
Khon Kaen University
Khon Kaen, Thailand
e-mail: somjit@kku.ac.th

***Abstract*—Buying behavior is considered a critical role in organizations' marketing management. Analyzing accurate buying behavior helps entrepreneurs to adopt production plan and marketing effectively. This research proposes an enhanced approach to analyzing buying behavior of targeted customers to acquire their high dimensional data on buying product patterns. The method is divided into three stages: First stage is clustering purchase of products based on type of customers by using K-Means and selecting the appropriate clusters by using Elbow method. The outcome from this stage was the similar purchased product items, which were individually categorized; in the second stage analyzing buying behavior using Apriori Algorithm. Then in ARA-1 we specify threshold 2 values. They are support and confidence not less than 10% and 70%, respectively and ARA-2 is adding the Lift values to the threshold not less than 1. The outcomes from these stages were buying patterns of the individual group. And the third stage is the accuracy of buying product patterns were evaluated by experts. In this study, the finding was trialed using purchasing data from a retail store in Thailand, revealing that the proposed method was likely to effectively analyze buying behavior with dimensional data. The accuracy of buying behavior analysis ARA-2 was higher than 88% and higher than ARA-1 38%. Interestingly, the approach revealed new buying behavior not previously reported.**

Keywords—Data mining; Clustering; Association rule; Behavior analysis

I. INTRODUCTION

The analysis of customer purchasing behavior for products or services has played a significant role in conducting the business particularly in marketing and product planning to response customer needs properly and effectively. However, product buying behavior especially commodity products complicated with diverse products and mass transaction lists, results in high dimensional data, thereby decreasing accuracy of the data on individual customers' buying products behavior.

Ilung Pranatac and Geoff Skinner [1] analyzed consumer buying behavior on 6 types of products among Portuguese wholesalers on annual expenditures by using K-Means. According to this study, the method could divide customers into three clusters, whose their average expenditure accounted for $90,000, $50,000 and $20,000, respectively. All these three clusters showed the similar buying behavior - selecting the best fresh food. Zhang Yihua's research [2] provided analysis to introduce the best communication phone plan for customers based on usability of customers. The researcher performed customer clustering by K-Means and analyzing the association between customers and communication phone plan by Decision Tree. It indicated that the proposed method could divide customer clusters into six groups and obtain the appropriate communication phone plan based on the customer characteristics.

Given these research approaches, it might be interesting to improve association rules as summarized and we propose a new solution as follow. Given problems concerning the customer buying behavior, we propose three processes. They are clustering the product by using K-Means process, analyzing association of buying behavior patterns of individual cluster by proposing an enhanced method and then compare it with the traditional method; and precision and accuracy in earlier analyzing association of buying behavior were evaluated by experts. Given the problems concerning high dimensional data of buying products, which results lowered accuracy of association rule, we added Lift threshold into the traditional threshold method. This paper divided into five sections. Section II concerns review of the theory and related work. Section III proposes the enhanced method. Experiment is discussed in section IV and concluded in section V.

II. BACKGROUND AND RELATED WORK

A. Data Mining

Data Mining [3] is the process of machine learning and statics in order to screen hidden knowledge in data warehouse, comprising 2 methods. First, data mining is achieved by creating (1) model through learning data of clusters previously specialize or predefined or through supervised learning and using the model to predict future actions. Second, data mining is achieved by unsupervised learning model. This model is used to screen data clusters whose answers or targets are not previously specified nor predicted. Both approaches rely on different technique and special characteristics as shown in Fig. 1 and Table 1.

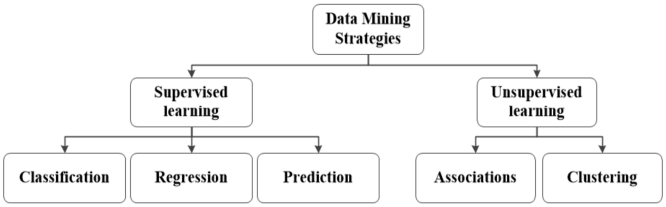


Fig. 1. Data Mining

TABLE I. SHOWS THE CHARACTER OF DATA MINING METHOD

Methods	Characteristics
Supervised Learning	• Learning the previous and current data to improve a model for further forecast • Data training requiring the coverage of the target data for effective performance
Unsupervised Learning	• Creating a model for data explanation • Due to absence of data training for machine learning, a sophisticated process of analysis is highly required.

B. Clustering

Clustering segmentation is to retain similar data in a cluster without learning from specific data. The characteristics used to cluster is to measure a distance between data; that is, any of 2 data with similar characters and short distance are retained in the same clusters. The accepted cluster should comprise 2 characters: The data within the same cluster should have a minimized intra-cluster distance and the data in different cluster should have different maximized inter-cluster distance.

Clustering comprises 2 methods [4]. Partitional clustering is a method of clustering according to the determined data and the number of required cluster, thus any of particular data would be categorized into a distinct particular group. Hierachical clustering is the method based on determined characteristics. As any of data have subset characteristics of another data, the divided cluster would be related to a large and small cluster in hierarchy. The number of cluster relies on the level of the relationship subject to divide.

C. K-means Clustering Algorithm

K-Mean is a non-hierarchical clustering method appropriately used for large data with high dimensional data and with few clusters. Prakash Singh and Aarohi Surya [5] compared the cluster performance of 9 algorithms: K-Means, Hierarchal clustering (birch), EM, MTree, Farthest first, canopy, LVQ, Cascading K-Mean and DBscan, suggesting that K-Means have the high performance appropriate for massive data clustering. [6] Moreover G.Jenitha and V.Vennila [7] suggested that it takes less time to create models when comparing to other algorithms. K-Means has following algorithm:

Specifying the desire number of cluster (specify by K) then select the centroid point of each cluster by randomize from all data after that calculate the distance between each data with each centroid point shown in equation 1 then cluster new data by considering between the distances of data with centroid point. And finally recalculate the distance between each data with centroid point.

$$d(x_i, x_j) = \sqrt{(|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots |x_{in} - x_{jn}|^2} \quad (1)$$

Whereas

$d(x_i, x_j)$= The distance between the data $x_i$and x_j

x_{in}= Attribute n of the data i

x_{jn}= Attribute n of the data j

n = Number of attributes

Clustering performance was evaluated by considering Sum of Squared Error: SSE, which is a measurement of the distance of SSE among the nearest centroid point, as shown in Equation 2.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(m_i, x) \quad (2)$$

Whereas

X = member data of the group C_i

m_i = Centroid point of the groupC_i

Elbow is the method used to select the best K for K-Means algorithms. [8] The differences of the SSE value of each K was measured for the considered K, in which the best K is Ki, where SSE value had no a significant difference to that of Ki+ 1 as shown in fig. 2. The values at K2 to K11 changing is shown in the graph, suggesting that, the SSE value of K5 and K6 has no a significant difference, thus the appropriate K is K5.

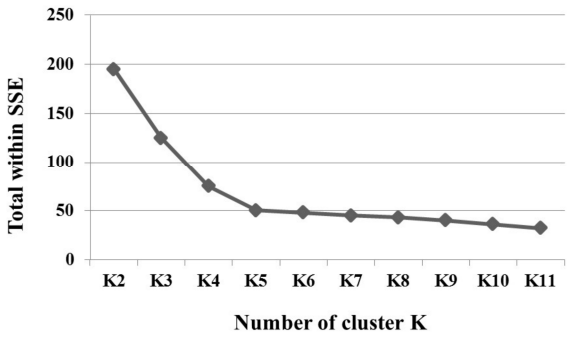


Fig. 2. Elbow Method

D. Association Rule

Abhijit Raorane and R.V.Kulkarni [9] said customer behavior is learning behavior between cluster or organization which have associate with product to response the customer and social requirements. Association Mining Rule is the technique used to be analysis customer buying behavior. Agrawal, Rand et al. [10] defines Association rule as follows:

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n binary attributes called items. Let $D = \{i_1, i_2, \ldots, i_m\}$ be a set of transactions called the database. Each transaction in *D* has a unique transaction ID (TID) and contains a subset of the items in *I*. A rule is defined in the form: X→Y, where X, Y subset of *I*. A rule is defined only between a set and a single item, X→ i_j for $i_j$is a member

of *I*. Every rule is composed by two different sets of items, also known as item-sets, X and Y, where X is called antecedent or left-hand-side (LHS) and Y consequent or right-hand-side (RHS). Normally consider selected rule base on minimum thresholds 2 values are support and confidence. Support is an indication of how frequently the item-set appears in the dataset. Let X is an item-set, X→Y an association rule and T a set of transactions of a given database. The support of X (support (X)) with respect to T is defined as the proportion of transaction in the dataset which contains the item-set X.

$$Support\ (X) = \frac{|\{t \in T; X \subseteq t\}|}{|T|} \quad (3)$$

Confidence is an indication of how often the rule has been found to be true. The confidence value of a rule, X→Y, with respect to a set of transactionsT, is the proportion of the transactions that contains X which also contains Y.

$$Confidence(X \rightarrow Y) = \frac{Support(X,Y)}{Support(X)} \quad (4)$$

Whereas

Support(X, Y) is the support value of item-set X and item-set Y that occurrence at the same time.

Support(X) is the support value which representing the occurrence frequency of item-set X in transaction D.

Beside these 2 thresholds, incase if have high dimensional data for example having a lot number of attribute and number of transaction made the rule have low accuracy. Therefore it should have Lift value indicator and choose the Lift value greater than 1 to ensure that the chosen rule, is strong rule. It means when item-Set X occur then item-set Y will definitely occur according to determine support and confident value. Lift value was calculated in equation 5.

$$Lift(X \rightarrow Y) = \frac{Support(X,Y)}{Support(X) \times Support(Y)} \quad (5)$$

E. Apriori Algorithm

Apriori Algorithm is an algorithm used to analyze association rule.[11,12] An association rule can be divided into 2 steps: Analyze frequent item-set in the transaction should have frequency not less than threshold support and follow the analyze frequent item-set result to crate association rule which corresponds with threshold confidence as shown process in fig 3.

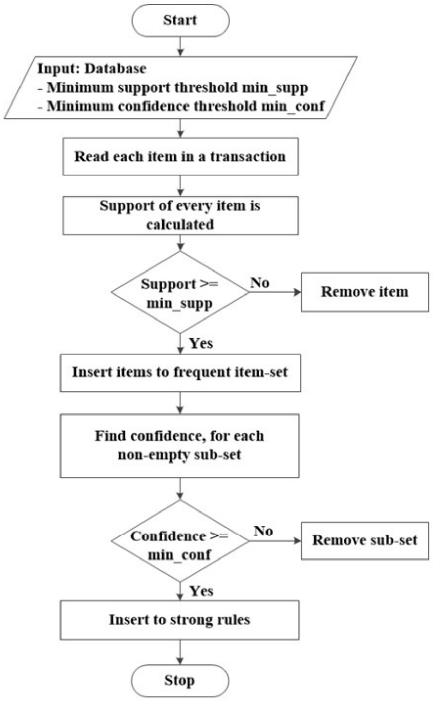


Fig. 3. Apriori Algorithm

F. Related Work

Ilung Pranatac and Geoff Skinner [1] conducted their study by clustering customer by using K-Means algorithm. They selected the K value by using Elbow method to obtain the suitable customer cluster. Rokhmatu lInsani and Hira Laksmiwati Soemitro [13] analyzed and compared K-Means and Kohonen SOM (sef-organizing map) for clustering customer. In this study, the researchers assigned K values by themselves and analyzed association rules of the communication industry promotion package by using Apriori algorithm, which determined support and confidence values at 40% and 50% respectively. This may obtain an association rule but lack accuracy. Therefore, this research enhanced the approach to selecting appropriates K for K-Means through the addition of Lift value to increase the high performance of association rule.

III. PROPOSED APPROACH

This research proposes an enhanced method for analyzing high dimensional data and several transactions of purchasing behavior. The study wherewith was divided into 3 processes: Firstly, Clustering the buying product list as customer type, second analyzing association rules of customer buying products and evaluating the obtained association rules by an expert. The conceptual framework is shown in fig. 4.

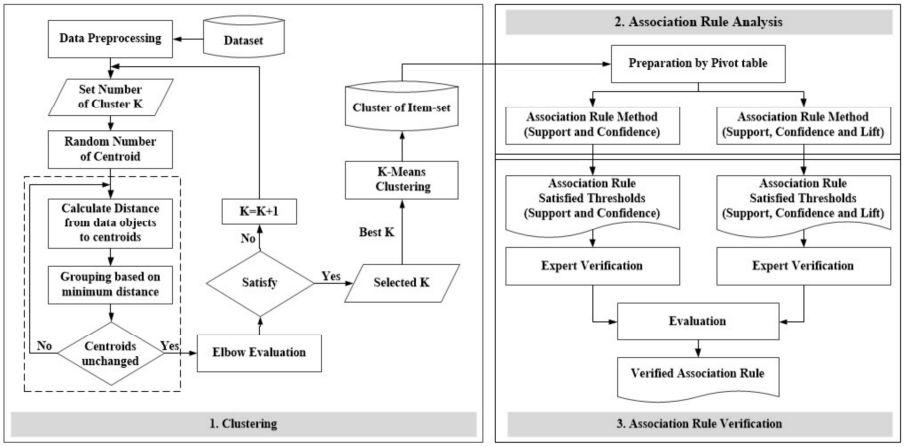


Fig. 4. Conceptual FrameworkClustering

According to in the preliminary experiment result in which an analysis association customer's buying product was performed without clustering gain, the result was found the unsupported thresholds support and confidence and that association rules for customer was found weak. Therefore, in the early first step, this research clustered types of customers by using K-Means algorithm and subsequently selected the appropriate number of cluster (K) by using Elbow method as following steps.

1) Data preparation step: customer buying product transaction was selected based on important attributes for analysis, for instance bill no., type of customer, product lists, etc.

2) K selection step: use the data obtained from the step one to cluster by K-Means, initiated by K=2 to n and select the best K from each SSE value by using Elbow method.

3) Clustering step: select the best K from 2) and data from 1) to cluster the data based on the types of customers by using K-Means. The obtained results were clustered product list equal to the number of K. Each cluster may contain more than one type of customer.

A. Association Rule Analysis

Transaction of each clustering was analyzed for association rules or customer buying behavior by 2 methods; 1) Association Rule Analysis -1 (ARA-1): analyzing association rules by using traditional thresholds; support and confidence by determining value not less than 10% and 70% respectively, and 2) Association Rule Analysis-2 (ARA-2): adding additional threshold, using Lift greater than 1.00, to ARA-1as following steps.

1) Data preparation step: transforming new transaction of each cluster by changing each transaction product list from row to attributes where each row refers to bill no and column contains attributes of product list.

2) Association Rule analysis procedure: conducting the association rules by using Apriori Algorithm using 2 methods.

Association Rule Analysis by ARA-1 method, which is evaluated by basic threshold and appropriate for general high dimensional data. And Association Rule Analysis by ARA-2 method, which uses Lift value to evaluate association rules added from ARA-1, thus creating the strong association rules appropriate for high dimensional data such as product lists.

B. Association Rule Verification

Buying product association rule of each cluster was verified by an expert who individually evaluated the previous association rule ARA-1 and ARA-2 in order to compare the accuracy.

IV. Experimental Results and Discussion

A. Dataset

Experimental data were obtained from a store at Khon Kaen Thailand totally 22,450 instants. The components and attributes subject to analysis included customer type, product type, and bill no. and product list as shown below.

1) Have 3 customer types; 101 Food Service stores, 4 Industries and 15 wholesales

2) There were 4 attributes: customer type, product type, bill no., and product list.

3) There were 199 products lists such as spicy wings, blood, egg no.3 etc. and 17 products type such as processed eggs, grade B processed food, etc.

B. Clustering

Clustering analysis result of customer using K-Means and selecting the best K value by Elbow method whereof considering SSE value is shown on Fig. 5, suggesting that the appropriate K is equal to 5. The clustered customers by K-Means and used K=5 to cluster are as shown in table 2.

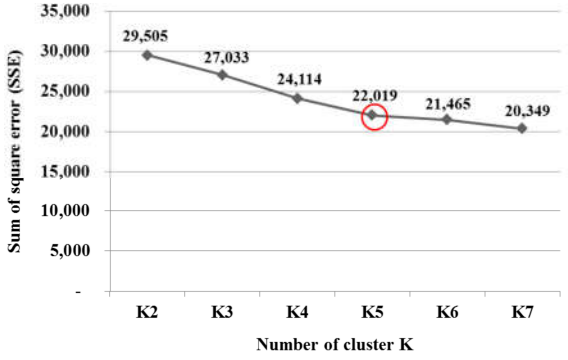


Fig. 5. Clustering analysis result of customer.

TABLE II. SHOWS THE ANALYSIS OF ASSOCIATION RULES OF PRODUCT PURCHASED BY INDIVIDUAL CUSTOMER GROUPS

Cluster	Customer Type	Product Type	Product List	Bill No.	Instance
1	Wholesale 55.09%,Food Service 44.78%,Industry 0.13%	Egg 99.53%, Cooked Product 0.17%, Marinade Product 0.14%, Processed Food Grade B 0.08%, Processedan Egg 0.04%, Butcher Seperated Pork 0.02%, Duck (Hymeat) 0.02%,	20	1,529	4,815
2	Food Service 93.01%, Wholesale 6.99%, Industry 0%	Branded Egg 100%	12	1,889	1,962
3	Food Service 99.28%, Wholesale 0.49%, Industry 0.23%	Chicken Parts BF/BFI 100%	26	1,594	2,526
4	Food Service 99.32%, Wholesale 0.48%, Industry 0.2%	Processed Food Grade A 52.89%, Duck (Hymeat) 13.96%, Butcher Seperated Pork 13.69%, Cooked Product 13.66%, Native chicken (Marinade) 1.97%, Meat Ball Premium 1.13%,Processed Food Grade B 1.09%, Seperated Pork BSM 0.13%, Native chicken 0.73%, Seasoning & Sauce 0.32%, Marinade Product 0.27%, Concentrated Soup 0.08%, Processed Egg 0.08%	88	2,162	3,397
5	Food Service 97.83%, Wholesale 2.04%, Industry 0.13%	Seperated Pork (Local) 100%	67	5,573	9,750
Total			**213**	**12,747**	**22,450**

Food service constitutes the majority of customers except the 1st cluster, which has same ratio with wholesale. Each cluster has different outstanding duct list and Cluster 4 has diverse product items, suggesting that in this cluster customers buy different goods.

C. Association Rule and Verification

Buying data of each cluster was analyzed for an association rule of buying products using 2 methods: Association Rule Analysis -1 (ARA-1) and Association Rule Analysis -2 (ARA-2). The experiment suggests that customers' buying data contained high dimensional data, thus specifying high value of Support and Confidence as explained above, thus no association rules were

The association rule created by both methods provided the precision and accuracy as shown on Table 3, indicating that ARA-2 had accuracy 88% and ARA-1 accuracy 64%. This indicates that ARA-2 had higher value than ARA-1 38% and the amount of associate rule in ARA-2 had less than ARA-1. This suggests that the association rules were filtered, thus remaining mere strong rules.

Cluster 4 shows no associate rules neither ARA-1 nor ARA-2 method. It indicates that this cluster consists of product items up to 88 items from overall 199 items, accounting for 44% of product item and most of which fall in by food service customers. This type of customer type was also shown in buying products activity in other clusters and which occur association rule. That means cluster 4 food service customer also buy other product items, which is considered not main product items.

Moreover, according to the analysis of the association rules by ARA-2 method, that the new association rule in cluster 5 which had not been acknowledged by the expert. This suggests that food service customer buying products and having new association rule, which is "Pork Leg H-1 paired with Pork Chop (pack)", "Pork Leg H/C pair Pork Chop (pack)" and "Pork mince B1 kg pair with Pork Chop (pack)". Given the early 2 association rule, food service customers are unlikely to buy pork leg paired with pork chop; however, the association rule was revealed. This is maybe because some types of the food service customers may run stewed pork leg restaurant and steak house.Fig. 6 and 7 indicates that ARA-2 has higher precision and accuracy than ARA-1 88%, which is the highest of all clusters.

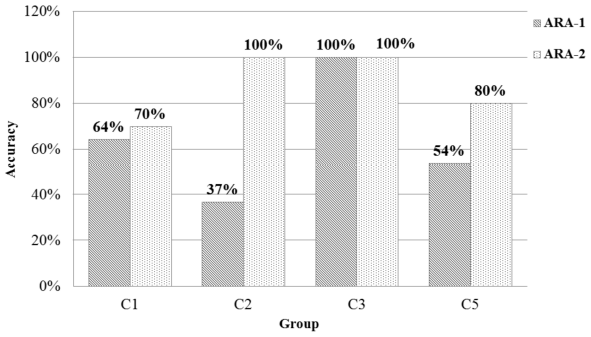


Fig. 6. Comparing performance precision and accuracy of association rule between ARA-1 and ARA-2 in each cluster

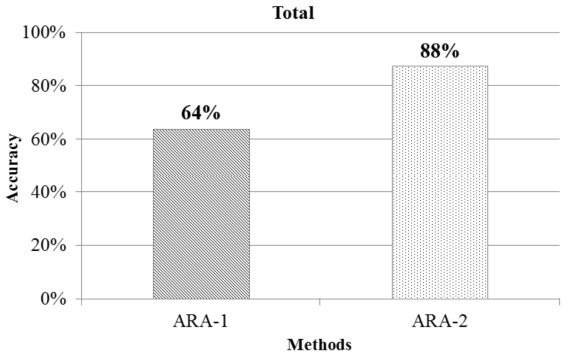


Fig. 7. Comparing performance precision and accuracy of association rule between ARA-1 and ARA-2

TABLE III. CLUSTERING ANALYSIS RESULT OF CUSTOMER BY K-MEANS

Cluster	Association Rule Analysis-1 (ARA-1)				Association Rule Analysis-2 (ARA-2)				
	Rule	*Support*	*Confidence*	*Expert Verification*	*Rule*	*Support*	*Confidence*	*Lift*	*Expert Verification*
1	7	35%	81%	64%	5	40%	77%	1.23	70%
2	3	10%	98%	37%	1	10%	99%	1.01	100%
3	4	10%	73%	100%	4	10%	73%	1.67	100%
5	5	10%	97%	54%	3	10%	98%	1.01	80%
Average				**64%**	**Average**				**88%**

Note: association rule were not found, thus eliminating evaluation in cluster 4

V. CONCLUSIONS

According to the analysis of buying products with high dimensional data by clustering customers' buying product using K-Means, selecting the best K by using Elbow method, it makes it likely to reveals association rules with high performance. Meanwhile, creating association rules by basic threshold Support and Confidence produces the weak association rules. However, by adding threshold by Lift value, it provides strong precision and accurate association rules, which was verified for the precision and accuracy 38%. Interesting, a new association rule was also reported.

However, the future research would make this research more complete by increasing process velocity, especially association rules for mass data. Furthermore, reducing frequent item -set may help to evaluate velocity.

ACKNOWLEDGMENT

Thank for Betagro fresh food store Klang muang Khon Kaen province, Thailand for the data ownership in this research.

REFERENCES

[1] IlungPranata, and Geoff Skinner, "Segmenting and targeting customers through clusters selection & analysis", IEEE International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2015.

[2] ZHANG Yihua, "Vip Customer Segmentation Based on Data Mining in Mobile-communications Industry", 5th International Conference on Computer Science & Education, 2010.

[3] Mark H, lan W, and Eibe F, "Data Mining Practical Machine Leaning Tools and Techniques", Morgan Kaufmann Publishers, 2011.

[4] AmitSaxena, Mukesh Prasad, Akshansh Gupta, NehaBharill, Om Prakash Patel, ArunaTiwari, MengJooEr, Weiping Ding, and Chin-Teng Lin, "A review of clustering techniques and developments. Neurocomputing", V267, 2017

[5] Prakash Singh, and Aarohi Surya, "Performance Analysis of Clustering Algorithms in Data Mining in weka", International Journal of Advances in Engineering & Technology, Vol. 7, Issue 6, pp. 1866-1873, 2015.

[6] KMArchana Patel, and PrateekThakral, "The Best Clustering Algorithms in Data Mining", IEEE International Conference on Communication and Signal Processing, April 6-8,(2016).

[7] G.Jenitha, and V.Vennila, "Comparing the Partitional and Density Based Clustering Algorithms by Using Weka Tool", 2nd IEEE International Conference on Current Trends in Engineering and Technology (ICCTET), 2014.

[8] Trupti M. Kodinariya, and Dr. Prashant R. Makwana, "Review on determining number of Cluster in K-Means Clustering", International Journal of Advance Research in Computer Science and Management Studies, V1, Issue 6, pg. 90-95. –10, 2013.

[9] AbhijitRaorane, and R.V.Kulkarni, "Data Mining Techniques: A Source for Consumer Behavior Analysis", Cornell University Library, .2011

[10] Rakesh Agrawal, Tomasz Imielinski, and Arun Swami, 1993 "Mining association rules between sets of items in large databases", Proceedings of the 1993ACM SIGMOD international conference on Management of data - SIGMOD '.93p..07doi: .170035.170072/10.1145ISBN .0897915925

[11] Alexander Setiawan, Gregorius Satia Budhi, Djoni Haryadi Setiabudi, Ricky Djunaidy, "Data Mining Applications for Sales Information System Using Market Basket Analysis on Stationery Company", IEEE International Conference on Soft Computing, Intelligent System and Information Technology, .2017

[12] Warnia Nengsih, "A Comparative Study on Market Basket Analysis and Apriori Association Technique", IEEE 3rd International Conference on Information and Communication Technology (ICoICT), 2015.

[13] Rokhmatu lInsani, and Hira Laksmiwati Soemitro, "Data Mining for Marketing in Telecommunication Industry", IEEE Region 10 Symposium (TENSYMP),2016.

Face Detection Using Haar Cascade in Difference Illumination

Diajeng Tyas Purwa Hapsari
Magister Teknik Informatika
Universitas Dian Nuswantoro
Semarang, Indonesia
ajenghapsa@gmail.com

Cindykia Gusti Berliana
Magister Teknik Informatika
Universitas Dian Nuswantoro
Semarang, Indonesia
cindykiagustiberliana@gmail.com

Putri Winda
Magister Teknik Informatika
Universitas Dian Nuswantoro
Semarang, Indonesia
putrwinda444@gmail.com

Dr. M.Arief Soeleman
Magister Teknik Informatika
Universitas Dian Nuswantoro
Semarang, Indonesia
arief2802@dsn.dinus.ac.id

***Abstract*-In this era of advanced technology, camera which has equipped with hi-tech features is no longer a new invention. It has the capability to recognize each of face parts as an object, the major development of biometric system. Face detection is still enthusiastically developed by identifying an individual object in digital image, analyzing and comparing its pattern. As we know that face is a real object which reflects self-identity and differentiates mankind from one another. Face detection can be used to find and index pictures and videos with background, size and position. There are some methods of face detection. One of these is haar cascade. It allows the system to recognize people's face with varied illumination.**

Keyword-face detection, haar cascade, TF-IDF

I. INTRODUCTION

Mankind is unique creature with face characteristics besides their mindset and behaviour. Human recognition with picture as an image object is an interesting study to be investigated further. Technology and its rapid shift is closely related to computer image processing. It starts with analog and now has shifted to digital technology.

Image processing is a system that processes the image-based input and produces the image-based output. The system produces a better quality image as the user's intention. In brief, it is aimed to improve the result of the image. [1] Human face is a dynamic object with varied shape and color [2], yet that causes more problems compared to another image object [3]. Object detection discover presence of an object within an image or video frame[4]. In this case, the labelled-class images are positive (the face is detected) and negative (the face is not detected) The purpose of face detection is to determine the existence of face in a picture and if it is positive, the location and size of each of faces will be detected and spotted as face. For most of computer vision observer and researcher, face detection system is a promising opportunity both commercially and governmentally that has to be improved.

The use of feature haar significally enriches simple features [5] and can be processed efficiently [6]. Haar feature is association function of two dimension (2D) completed haar that codes the image of local object [7].

II. REVIEW OF RELATED LITERATURE HAAR FEATURE

A. RELATED REVIEW

Kritika Shrivastava proposed an automated attendance system with recorded output of classroom session by which the lecturer or faculty can record student's attendance with minimum human interface (less human error) or student proxy attendance. This system increases its efficiency and reliability toward recognition of faces of the students to provide a secure system.

Om Patil proposed a gesture recognition with Haar-like features and the AdaBoost learning algorithm. The Haarlike features can effectively describe the hand posture pattern with the analytically computed. The AdaBoost learning algorithm can greatly speed up performance and construct a strong classifier by combining many weak classifiers. This structure can achieve satisfactory real-time performance and high classification accuracy. This paper tries to give solution to the communication problem faced by many people across the globe. This system may use most of the processing power of the processor.

K.J. Raman proposed In the research area of computer vision and image processing, face recognition is a challenging portion that has received a great deal of consideration and demands over the last few years because of its applications in various domains. This research can be expanded by using other features such as sensor motion in different aspects to reduce road acciden.

B. RELATED WORK

In this paper I want to analyze system performance using the Haar cascade method for face detection in various lighting conditions of the sun and lights. Face detection is the segmentation process of the face of the background image with the varied illumination as one of the comparative factors. The principle of the process is the parameter of the image is detected whether it has face characteristic or not. If it has the separation stage of face and background, the image will be processed.

For testing this algorythm an image input will be pre-processed before the face detection process.

Pre-process in face detection process includes changing in image size and color to grayscale. If in this process face image is found, then the system will determine the best version of the face image. Object detection will improve the image orientation until the best version of the face image is taken as the improvement of the image orientation by rotating to the right direction. After that process we used confussion matrix to measure the level of accuracy.

C. PROPOSED METHOD

1. HAAR FEATURE

Haar feature is a wavelet-based feature that decomposes image. The word "haar" refers to mathematics function that has rectangle shape [8]. Wavelet Haar is a single rectangle wave (one high interval and one low interval). For two dimension (2D), it has one light side and one dark side. The function of cascade classification is to combine more feature efficiently.

In the beginning, the image process of haar is only based on RGB value of each pixel, then process the image in rectangle shapes with some pixel in every shape. Each of the shapes is processed and the limit level (threshold) is occurred after that which shows dark and light area. The formula of haar feature is the average value of the result is above the threshold, it means the haar feature exists.

2. GRAYSCALE IMAGE

One of the pre-processes in face detection of mankind object is the change of the real image room into grayscale. That is occurred since generally facial object in grayscale image has consistent pattern such as eyes color is darker than cheeks or nose color. Here is the formula to turn color (RGB) into grayscale [9]

Grayscale = α*RED + β*GREEN+ γ*BLUE

The image that is produced in this stage later will be processed in characteristic extraction stage of the image and face recognition.

3.. Confussion Matrix

Confussion matrices as a means of counting the cases in hypothesis agress or disagree with the teacher signal. The first measure one is usually intersted in is "precisenss". Preciseness means not making a wrong predictionin the first place[10].

4. Cascade Adaboost Classifier

Cascade AdaBoost classifier is a model of classification with some levels. Each of the classifications is formed by adaboost algorithm. Principally whether the framework of classification model is categorized as good or not depends on three things as follows, number of classification level, number of feature, and threshold value at each classification level, but all of them are obtained based on input parameter which are the number of feature, detection rate and false alarm rate. Furthermore, the contents of model cascade adaboost classifier are some feature harlike and threshold value. A sample is valued positive as object (the feature value is higher from threshold) if it passes on all classification levels, otherwise it is negative (the sample is not valued as an object). Whether or not this classification method cascade adaboost in term of reducing computation when this method is used for face detection in an image.

III. ANALYSIS AND RESULT

A. DATA SET

In this paper we use webcam to crawl the data, which is take size

B. FACE DETECTION AND TESTING

Face detection is segmentation process of face area against the background of the input image with the varied illumination as one of the comparative factors. The principle of the system of this process is the parameter of the image is detected whether it has the face characteristic or not. If it has the separation stage of face and background, image will be processed. Here is the diagram of face detection:

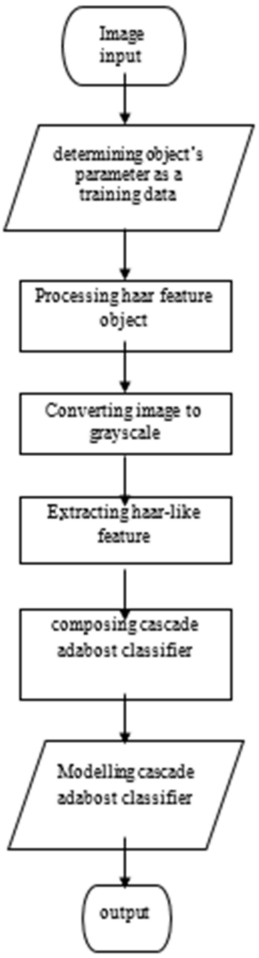


Figure 1. Flow diagram of pre-face detection

Figure 1 is flow diagram of a training process that shows a classification model, while

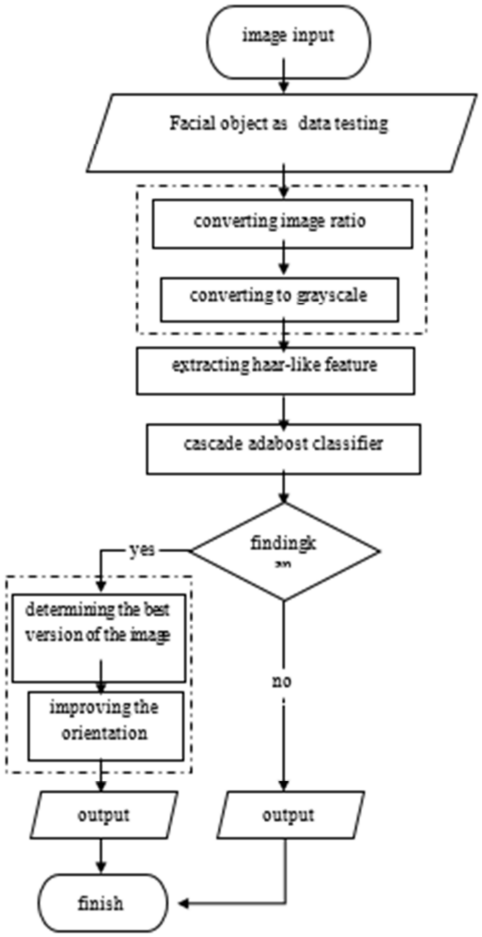


Figure 2. Flow diagram if face testing

figure 2 is the testing. In testing stage, an image input will be pre-processed before the face detection process. Pre-process in face detection process includes changing in image size and color to grayscale. If in this process face image is found, then the system will determine the best version of the face image. Object detection will improve the image orientation until the best version of the face image is taken as the improvement of the image orientation by rotating to the right direction.

FACE TESTING ON BRIGHT, MEDIUM AND DIM CONDITIONS

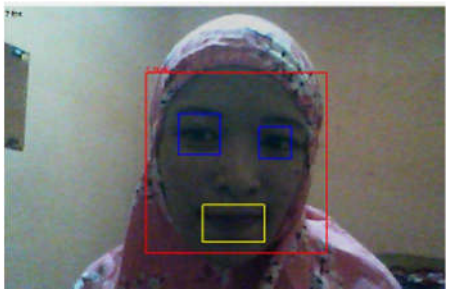


Figure 3. Face detection on dim condition

In figure 3, the eyes are detected as the darkest part of the object

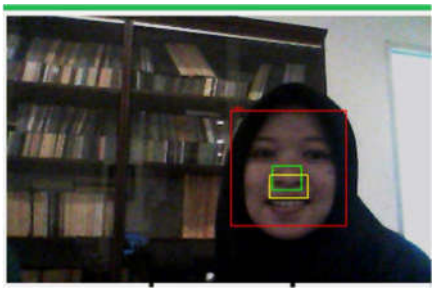


Figure 4. Face detection on medium light condition

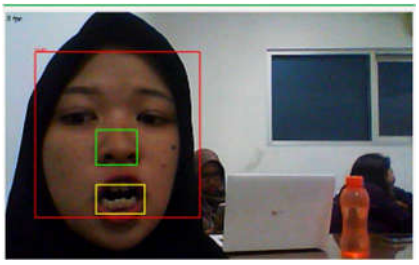


Figure 5. Face detection on bright condition

In figure 3 the value is 7fps, while figure 4 and 5 each has value 3fps

Detection still gets its best performance since it uses distance that provides optimal value that are at 30 cm to 70 cm.

Table 1. Lamp illumination

Object	TF							IDF
	D1	D2	D3	D4	D5	D6	D7	
Eyes	1	1	1	1	1	0	0	0,54
Nose	1	1	1	1	1	0	0	0,24
Lips	1	1	1	1	1	0	0	0,36
Face	1	1	1	0	1	1	1	0
W								
D1	D2	D3	D4	D5	D6	D7		
1,54	1,54	0	0	0	0	0		
1,54	1,54	1,5	1,5	0	0	0		
1,54	1,54	1,5	0	0	0	0		
1,54	1,54	1,5	1,5	1,5	1,5	1,5		
Weight								
6,16	6,16	4,5	3,0	1,5	1,5	0,5		

In Table 1 the last experiment mengalai weighting decline as more away from the webcam face objects thereby reducing the accuracy of detection.

Table 2. Solar illumination

Object	TF							IDF
	D1	D2	D3	D4	D5	D6	D7	
Eye	1	1	0	0	0	0	0	1,54
Nose	1	1	1	1	1	1	0	1,06
Lip	1	1	1	1	1	1	0	1,06
Face	1	1	1	0	1	1	1	0
W								
D1	D2	D3	D4	D5	D6	D7		
1,54	1,54	0	0	0	0	0		
1,06	1,06	1,06	1,06	1,06	1,06	0		
1,06	1,06	1,06	1,06	1,06	1,06	0		
0	0	0	0	0	0	0		
Weight								
3,66	3,66	2,12	2,12	2,12	2,12	0		

In table 2 the last experiment weighting has decreased as more faces from webcam objects away so as to reduce the detection accuracy in the detection of this solar light on face detection performance decreases more than detection using light bulbs.

From the test above, the result is that face, nose and eyes detection got optimal performance from the lamp illumination, depending on the light that captured by the camera. Sufficient light captured by the camera will allow the system to identify eyes, nose and mouth well. The difference is very contrast when tested on sufficient light.

IV. CONCLUSION AND SUGGESTION

Dim light affects the system performance, therefore it cannot detect the image completely. On

average, the detection speed of the system on moving object is good in every condition.
When the result is calculated with TF-IDF, the highest value is when the testing used the light from the lamp.
On TF IDF weighting 62.7% for lamps and 37.2% of the Sun's light. Making light of a lamp which could prove that face detection can work well if using light bulbs.
Good illumination will help and improve face detection in order to get good value and performance with TF-IDF.

REFERENCES

1. Pulung Nurtantio Andono, Muljono, T.Sutojo, "Digital", 2017
2. Nidhi, "Image Processing and Object Detection", Dept. of Computer Applications, NIT, Kurukshetra, Haryana, 1(9): 396-399, 2015.
3. Muller, N., Magaia, L. and Herbs B.M. Singular value decomposition, eigenfaces, and 3D reconstructions. *SIAM Review*, Vol. 46 Issue 3, pp. 518–545. Dec. 2004.
4. Grevelink, Evelyn (Intel), "A Closer Look at Object, Recognation and Tracking,"2017, Available: https://software.intel.com/en-us/articles/a-closer-look-at-object-detection-recognition-and-tracking
5. Viola, P., Jones, M.J.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: IEEE CVPR (2001).
6. Lienhart, R., Kuranov, A. and Pisarevsky, V. Empirical Analysis of Detection Cascade of Boosted Classifiers for Rapid Object Detection. pp. 297–304. (2003).
7. R. Lienhart, J. Maydt, An extended set of Haar-like features for rapid object detection, in: ICIP '02: Proceedings of the International Conference on Image Processing, 2002, pp. 900- 903.
8. Bahram, D., Elias, D., Shahram, D. And Omid, M. Discrete Wavelet Transforms of Haar's Wavelet, Vol.3 , pp.247-251, 2014.
9. M. Borawski and D. Frejlichowski. 2012.An Algorithm for The Automatic Estimation Of Image Orientation. The Scientific Word Journal Machine Learning AndData Mining in Pattern Recognition. Springer Berlin Heidelberg, Vol.7376, pp. 336-344.
10. Muller ME, Relational Knowledge Discovery, United Kingdom: University Press, Cambridge,pp 62

Search Engine for Kids with Document Filtering and Ranking Using Naive Bayes Classifier

Junta Zeniarja
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
junta@dsn.dinus.ac.id

Ramadhan Rakhmat Sani
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
ramadhan_rs@dsn.dinus.ac.id

Ardytha Luthfiarta
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
ardytha.luthfiarta@dsn.dinus.ac.id

Heru Agus Susanto
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
heru.agus.santoso@dsn.dinus.ac.id

Erwin Yudi Hidayat
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
erwin@dsn.dinus.ac.id

Abu Salam
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
abu.salam@dsn.dinus.ac.id

Leonardus Irfan Bayu Mahendra
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
leonardusirfan@gmail.com

***Abstract*—Search Engines are one of the important facilities used to facilitate users in searching information especially information from web documents available in internet. Without the help of search engines, users need to know the address of the intended web page to open certain web pages. Almost everyone uses search engines to search for information on the internet, no exception children. This can be worrying especially since most search engines are designed by and for adults in search of information. The wide variety of information available on the internet ranging from safe to insecure to children led to the need for specialized search engines designed specifically for the search for information appropriate for children. In this search engine, using the concept of Information Retrieval with Document Filtering as a pornography media filtering, so it is safe for children to use. Document filtering uses the Naive Bayes algorithm for the learning process in document categorization, where the information accessible to children is categorized as "safe" and which can't be accessed by children categorized as "unsafe". It is hoped that with this search engine application, it can help the child in searching for information safely although not accompanied by his parents.**

Keywords—search engine; childrens; filtering; naive bayes

I. INTRODUCTION

We are in a period when information access is virtually unlimited. The most significant example of technology is the internet. Through the internet we can find various kinds of information ranging from simple information that can help in daily activities up to information about the latest things that happen in foreign countries. Among the many piles of information on the Internet, looking for information that we need in the manual is not possible to do. That's why we need a search engine facility. The search engine will find the Web documents containing the information we need through a simple keyword. The high demand for this search engine is shown by statistics that google today is one of the most widely used search engines in the world, processing over 40,000 queries every second [1].

With the ease of accessing the internet today, children are already using the internet to find the information they need every day. Studies of childtrends.com said that in 2013, 57 percentage from children aged 3-17 years old use the Internet in their homes [2]. Search engines certainly be one of the facilities used, but it is becoming alarming because most search engines available in designs for adults in search of information. Statistics from Guardchild.com say that 70% of children aged 7 to 18 years old inadvertently find things that intersect with adult content while searching the internet for their homework [3]. The development of search engines for children's information especially in Indonesia is important to protect children from unsafe information on the internet. The information is not safe as well as violence, racial issues, gambling, and buying and selling narcotics.

In determining whether a web page is safe to display for children's users, it can be done by analyzing the use of words in the use of the web. Some words can be categorized as insecure words for children such as "Alcohol" "Drugs" and "Gang". Analysis of the use of these words can be done by using a Naive Bayes algorithm.

Naive Bayes Machine Learning Algorithm is one of the most widely used algorithms and is considered as a simple but accurate method as a method in document classification. Examples of applications from the Naive Bayes method are text classification [4], email filtering [5], recommendation systems [6], and search engine for university [7], [8]. Naive Bayes uses probabilities approach in classifying a document so that it can be applied in the search engines to determine how large the probability of a web document safe for children. Therefore, in this study attempted to solve the problems on parents who are still difficulties in control and secure the use of the Internet by children because of the small number of web document search engine designed specifically to search for information by children. So that can help parents to control and secure the use of the internet by children with a search engine designed specifically to search by children. And also do search

engine development oriented to information search by children with good accuracy

.

II. SEARCH ENGINE

Search Engines is a form of application of Information Retrieval techniques for large-scale text collections [9], [10]. Search engines themselves can be implemented into various domains such as library catalogs [11], [12]; university searches [7], [8]; or web document search [13]. The performance of search engines can be measured, for example, from the accuracy or relevance of the documents provided, the speed of response, the rate of document indexing, as well as the ability to be expanded or expanded. Here's the architecture diagram of the search engine:

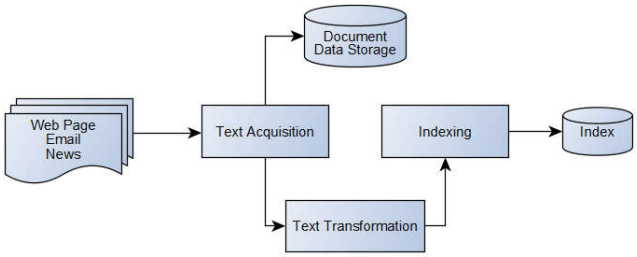


Fig. 1. Search Engine Architecture for System

In the fig. 1, the process of text acquisition on a search engine is usually done by a bot that is often also referred to as a crawler. The next step is text transformation, to preprocessing the text data that has been obtained. Indexing aims to speed up the process of searching documents based on keywords entered by the user as well as the basic ranking of documents that govern how the order of documents will be displayed as a result of the search to the user.

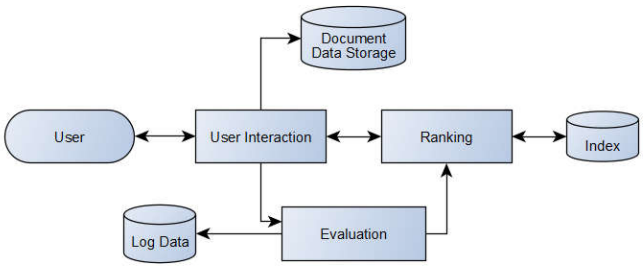


Fig. 2. Search Engine Architecture for User

In the fig. 2, User Interactions, search engines work based on the user's query input. The task of this step is how to translate the query from the user into a query that can be used to access the index database. In the ranking process, the order of the documents becomes very important for the search results that the number is very much. Based on feedback arising from interaction with users, evaluation results are used to correct the weaknesses of existing search engine models.

III. NAIVE BAYES CLASSIFIER

Naive Bayes is one of the classification method using probability approach [5], [6], [7], [11], [14], [15], [16]. Mathematically Bayes theorem can be written as follows:

$$P(h|D) = \frac{P(D|h) \times P(h)}{P(D)} \quad (1)$$

To determine the class of documents from some of the features of the documents obtained:

$$C_{NB} = \underset{c \in C}{\operatorname{argmax}} P(C_j) \prod_{x \in X} P(x \,|c) \quad (2)$$

IV. RESEARCH METHOD

A. *Framework*

The flow of this research process is described as follows:

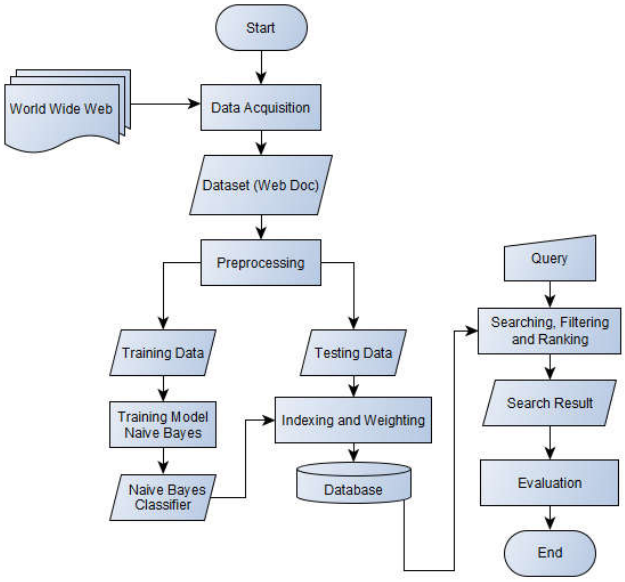


Fig. 3. Proposed model

1. Data Acquisition

The first stage in the research is to retrieve data, data that will be used for research, namely web documents that have been labeled as safe for children and unsafe for children. The data are divided into training data and test data with a ratio of 3: 1.

2. Preprocessing

The preprocessing process here is done to extract terms or important words that can be used as document features. The document preprocessing process is done with tokenization, stopword removal, and stemming.

3. Training Model with Naïve Bayes

Because Naive Bayes is a supervised learning method, before the classification is done it is necessary to do training to create a model as a basis for classification.

4. Indexing and Weighting

Index creation is made from test data. Index is created by calculating the number of occurrences of certain terms on web documents. As a core part of this children's search engine, a weighting measure is added to each document that determines the document's suitability to the search engine for children. This weight is calculated from the probability of document compatibility for children by applying the Naive Bayes probability model generated from the training phase in the previous phase. The probability weight generated by its value has a range of values from 0 to 1, with values close to 1 meaning that web documents are suitable to searching for children.

5. Searching, Filtering and Ranking

From the query entered by the user from the main page of the search engine, it will be matched with the index that has been stored in the database. First, specify which documents contain keyword queries. From the list of documents obtained, sort the search results based on the probability of matching obtained from the calculation with the Naive Bayes method. Display web documents obtained based on the highest probability results suitable for children. Filtering can be done by not including web pages with a probability of compatibility for children under a certain value.

6. Evaluation

The evaluation model process is carried out to calculate the accuracy of search results based on certain queries. The measurement process uses a Confusion Matrix diagram which is done by calculating Recall, Precision, and Accuracy.

B. Data Collection

Data is retrieved by selecting multiple pages that are identified as pages that are safe for child searches and pages identified as unsafe. The identification refers to the information displayed and the use of words on the web page. Here are some examples of words that refer to safe and unsafe for children:

Table 1. Example of Safe and Unsafe Words for Children.

Category	Words	Sentences
Safe	Education, smart, health, knowledge, science.	"We have to study in order to gain knowledge which is useful."
Unsafe	Drugs, gambling, murder, terrorism, crime.	"Fights made by some youths yesterday caused by drunkenness."

In collecting data, we initially creates a list of safe and unsafe web pages for children. Examples of web addresses used by us can be seen in the table 2.

Table 2. List of Web Pages.

No	Secured	Unsecured
1	Berani.id	Agenjudipoker.online
2	Kidnesia.com	Dewabet.asia
3	Farumanak.id	Kisahpenjudi.blogspot.com
4	Temankita.com	Rahasiaritualpesugihan.blogspot.co.id

To retrieve text from web pages, we uses the scrapping program made with the Python programming language results from the scrapping process is a text document containing a "bag of words" or a set of words that are used in the web page.

V. EXPERIMENT RESULTS

A. Interface Program

The search engine is implemented in the form of html pages as search pages for users. Here are the results of the user interface for the main search page and the search results page:

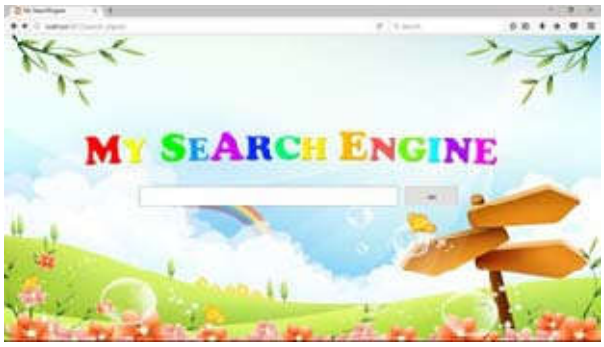


Fig 4. Interface Homepage of Search Engines

Fig 4 show the homepage interface of search engines displays a field box for search input with an appropriate background for children. Once the user inputs a keyword or query, then the search document will be displayed on the next page as shown in the fig 5 that the link the document resides on. The links shown are links from documents that are already classified as safe for children with the Naive Bayes Classifier algorithm.

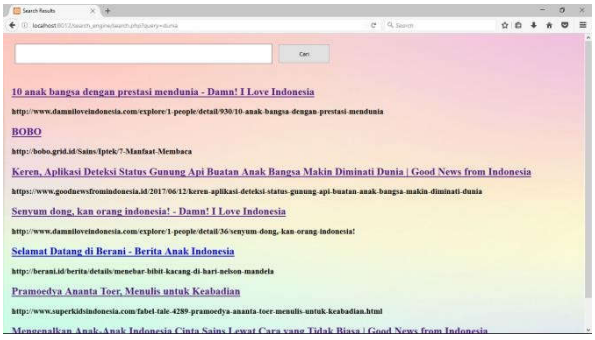


Fig 5. Interface of Search Engine Results

B. Result

The following classification results for all test documents with Naive Bayes classifier trained:

Table 3. Naive Bayes Classification Result

No	Class	Safe Probability	Unsafe Probability	Classification Result
1	Safe	3.81E-232	2.47E-238	Safe
2	Safe	4.80E-235	8.34E-278	Safe
3	Safe	8.86E-289	4.89E-307	Safe
4	Safe	5.81E-252	4.69E-268	Safe
5	Safe	1.14E-235	3.50E-244	Safe
6	Safe	1.19E-266	5.51E-261	Unsafe
7	Safe	1.19E-256	1.64E-264	Safe
8	Safe	3.90E-256	3.50E-264	Safe
9	Safe	6.54E-265	4.48E-278	Safe
10	Safe	5.57E-294	5.1876892813331 E-321	Safe
11	Safe	2.22E-271	6.88E-267	Unsafe
12	Safe	3.08E-270	3.44E-283	Safe
13	Safe	2.63E-254	7.06E-294	Safe
14	Safe	1.52E-282	1.39E-278	Unsafe
15	Safe	1.15E-250	6.35E-242	Unsafe
16	Safe	1.83E-225	5.95E-222	Unsafe
17	Safe	1.55E-249	1.51E-258	Safe
18	Safe	9.74E-268	5.13E-254	Unsafe
19	Safe	3.29E-235	4.62E-234	Unsafe
20	Safe	3.76E-288	2.89E-300	Safe
21	Unsafe	2.09E-278	9.91E-234	Unsafe
22	Unsafe	1.29E-256	2.68E-213	Unsafe
23	Unsafe	2.58E-258	6.46E-214	Unsafe
24	Unsafe	1.00E-285	1.02E-242	Unsafe
25	Unsafe	4.58E-232	2.74E-206	Unsafe
26	Unsafe	2.10E-237	1.48E-151	Unsafe
27	Unsafe	1.88E-303	1.55E-226	Unsafe
28	Unsafe	3.65E-304	5.36E-223	Unsafe
29	Unsafe	1.67E-294	1.78E-270	Unsafe
30	Unsafe	1.06E-243	4.16E-219	Unsafe
31	Unsafe	3.50E-233	1.47E-201	Unsafe
32	Unsafe	2.5414736822074 E-320	3.48E-256	Unsafe
33	Unsafe	3.46E-273	3.53E-209	Unsafe
34	Unsafe	4.77E-226	5.86E-179	Unsafe
35	Unsafe	1.83E-277	1.47E-22	Unsafe
36	Unsafe	5.8735177988513 E-311	1.39E-190	Unsafe
37	Unsafe	1.22E-237	3.17E-110	Unsafe
38	Unsafe	7.81E-107	1.90E-101	Unsafe
39	Unsafe	7.92E-265	2.39E-231	Unsafe
40	Unsafe	6.15E-235	1.03E-199	Unsafe

Based on the test results table above, Confusion Matrix for the classification results are discussed below:

$$Precision = \frac{TP}{TP + FP} = \frac{13}{13 + 0} = 1 \quad (3)$$

$$Recall = \frac{TP}{TP + FN} = \frac{13}{13 + 7} = 0{,}65 \quad (4)$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{13 + 20}{13 + 0 + 20 + 7} = 0{,}825 \quad (5)$$

Overall, the accuracy of this research done by Naive Bayes Classifier method in this study was 82.5%.

VI. Conclusion And Future Work

Based on the results of research conducted, it can be concluded that the Naive Bayes method can be used to perform classification and filtering of documents safe or unsafe for children. The resulting accuracy is quite high. Nevertheless, there are still some weaknesses that the author felt in the research results. One is the lack of relevance of the query with the document being generated due to the lack of weighting techniques and algorithms to measure the similarity of the document. Another disadvantage is that although high classification accuracy, cannot guarantee the classification for multimedia formats such as images, video, or sound. Query optimization is also required so that for the massive number of documents, search can still be done quickly.

In a subsequent study are expected to implement the weighting, so that the document can search more accurately to the query entered by the user. Coupled with the classification and filtering of multimedia formats such as images, video, and sound. Then query optimization, so that document search time can be shorter and faster for the number of documents that are massive (internet and world wide web).

References

[1] Internet Live Stats, “Google Search Statistics.” [Online]. Available: http://www.internetlivestats.com/google-search-statistics/. [Accessed: 20-Jul-2018].

[2] Child Trends DataBank, "Home computer access and internet use," 2015. [Online]. Available: https://www.childtrends.org/indicators/home-computer-access. [Accessed: 20-Jul-2018].

[3] Guard Child, "Internet Crime and Abuse Statistics." [Online]. Available: https://www.guardchild.com/statistics/. [Accessed: 20-Jul-2018].

[4] R. R. Sani, J. Zeniarja, and A. Luthfiarta, "Penerapan Algoritma K-Nearest Neighbor pada Information Retrieval dalam Penentuan Topik Referensi Tugas Akhir," *J. Appl. Intell. Syst.*, vol. 1, no. 2, pp. 123–133, 2016.

[5] M. Esmaeili, "An Anti-Spam System using Naive Bayes Method and Feature Selection Methods," *Int. J. Comput. Appl.*, vol. 165, no. 4, pp. 1–5, 2017.

[6] M. Ghazanfar and A. Prugel-Bennett, "An Improved Switching Hybrid Recommender System Using Naive Bayes Classifier and Collaborative Filtering," *Computer (Long. Beach. Calif).*, vol. I, pp. 493–502, 2010.

[7] J. Zeniarja, A. Luthfiartha, and C. Supriyanto, "Naive Bayes Classifier Based Geographic Information System for University Search Information," *J. Appl. Intell. Syst.*, vol. 2, no. 2, pp. 73–79, 2017.

[8] J. Zeniarja, A. Luthfiarta, and C. Supriyanto, "Aplikasi Pencarian Perguruan Tinggi Dengan Algoritma K- Nearest Neighbor Berbasis Information Retrieval Dan," in *Seminar Nasional Teknologi Informasi Dan Aplikasi Komputer (SINTAK)*, 2017, pp. 137–146.

[9] C. D. Manning and P. Raghavan, "An Introduction to Information Retrieval," 2009.

[10] W. B. Croft, D. Metzler, and T. Strohman, *Information retrieval in practice*. 2015.

[11] A. Setiawan, I. Fitri Astuti, and A. Harsa Kridalaksana, "Klasifikasi Dan Pencarian Buku Referensi Akademik Menggunakan Metode Naïve Bayes Classifier (Nbc) (Studi Kasus: Perpustakaan Daerah Provinsi Kalimantan Timur)," *J. Inform. Mulawarman*, vol. 10, no. 1, 2015.

[12] A. P. Kusuma and I. Srirahayu, "Antivirus: Jurnal Ilmiah dan Teknik Informatika, November 2016, Vol.10, No. 2 50," *J. Antivirus*, vol. 10, no. ISSN: 1978-5232, pp. 50–55, 2016.

[13] N. Juliasari and J. C. Sitompul, "Aplikasi Search Engine Dengan Metode Depth First Search (Dfs)," *BIT*, vol. 9, no. 1, pp. 9–12, 2012.

[14] D. Santoso, D. E. Ratnawati, and Indriati, "Perbandingan Kinerja Metode Naive Bayes, K-Nearest Neighbor, dan Metode Gabungan K-Means dan LVQ dalam Pengkategorian Buku Komputer Berbahasa Indonesia berdasarkan Judul dan Sinopsis," *Repos. J. Mhs. PTIIK UB*, vol. 4, no. 9, 2014.

[15] D. Sapkale and P. S. Nair, "An Improved Domain Classification of Google Search Results Using Naive Bayes Classifier," *Int. J. Eng. Sci. Comput.*, vol. 6, no. 7, 2016.

[16] A. Jain and J. Mandowara, "Text Classification by Combining Text Classifiers to Improve the Efficiency of Classification," *Int. J. Comput. Appl.*, vol. 6, no. 2, pp. 126–129, 2016.

Automatic Essay Grading System for Japanese Language Examination Using Winnowing Algorithm

Anak Agung Putri Ratna, Dyah Lalita Luhurkinanti, Ihsan Ibrahim, Diyanatul Husna, Prima Dewi Purnamasari
Department of Electrical Engineering
Faculty of Engineering, Universitas Indonesia
Depok, Indonesia
ratna@eng.ui.ac.id

***Abstract*—This paper explained a research that was conducted to find the best methodology for automatic essay grading. The Winnowing method was used in this paper with specific application for Japanese language exam. With Japanese language input, romanization was implemented to change the input that were in hiragana, katakana, or kanji into romaji. After the system was developed, it still needed to find the best parameters with the highest accuracy or closer even similar to the results from human rater. From testing the short essay Japanese language exam answers with 100% of score from human rater, several testing show that the best parameter for this system is with n=2, w=2, and p=2 and the average accuracy of total score gathered for all students is 86.86%. Since the average accuracy is above 80%, it could be concluded that winnowing algorithm is suitable for Japanese language examination term.**

Keywords—essay grading; fingerprint; hash; n-gram; winnowing

I. INTRODUCTION

Since 2007, Department of Electrical Engineering had developed an automatic grading system called SIMPLE-O [1]. This system is developed to ease the lecturer to grade essay, especially for Indonesian language short essay examination. Along with the increasing number of students, the number of documents is also increasing. The lecturer would be challenged to maintain its objectivity to scoring the exam when the exam answer also has a complexity to understand its text meaning. Then, this challenge would lead to exhausting process for the lecturer and lack of objectivity consistency. SIMPLE-O is built based on Latent Semantic Analysis (LSA) method for examining the real meaning of a context that was in form of sentence, paragraph, or a document. This method is not using sentences and language structures in its main process but using intense mathematical analysis process. But, as the purpose of this research mainly to develop the best system for daily use with the best accuracy, the system undergoes various modification with several different methods. Then, in this paper, this system was tried to implement Winnowing algorithm that based on fingerprinting process and the application of the developed system with that method is for Japanese language. After the system is developed with Winnowing algorithm, it still needed some tuning and configuration due to find the best running configuration for obtaining the best result when compared to human rater results.

This research briefly explained in four other sections. The concept of Winnowing algorithm and the other method of developing the system is explained in Section 2. Section 3 briefly explained the design of the system algorithm with Winnowing algorithm and its specific pre-processing for Japanese language application. Then, results and analysis of system testing in Japanese language essay grading are explained in the Section 4. And finally, the conclusion of this research is written in Section 5.

II. AUTOMATIC ESSAY GRADING WITH WINNOWING ALGORITHM

Winnowing is a fingerprint-based algorithm that was formed from a process called hashing. This fingerprint is numeric values which can be obtained by calculating the ASCII values of every character in n-gram. Fingerprint method is one of the most common techniques for comparing the similarity between texts or documents. It selects substrings that represent the text [2]. Winnowing algorithm itself solved the fingerprints selection's problem that happened in fingerprint algorithm by adding windowing process and selecting the fingerprint based on the hash on each window [3]. This algorithm can be used for automatic essay grading because it has some characteristics, such as whitespace intensity, noise suppression, and position independence [4]:

1) Whitespace intensity: the comparing process is not affected by whitespaces like space, capitalization, and punctuation.

2) Noise suppression: the used of the longer match to prevent any words that too short like a preposition.

3) Position independence: the comparing process is not affected by the arrangement of sentences.

Input for this algorithm are texts that will be processed to generate a group of minimum hash or fingerprint as the output. These fingerprint values are those that will be compared to one and another text.

Before processing is done by winnowing algorithms, first, it must go through the data pre-processing. It is intended to prepare the data, so the results are more accurate. Furthermore, there is a need for a method to process Japanese characters input. The winnowing algorithm has several processes, such as the

determination of the n-gram, hashing, windowing, and fingerprinting.

A. Pre-processing

Pre-processing is a process which is done before the data was processed using the winnowing algorithm. The functions of this process are to prepare data, simplify processes, and further increase the accuracy of the system. In this process, there is also the processing of the inputs that are still in the form of Japanese characters into romaji, the Romanized version of Japanese characters, so the ASCII values of that text can be searched. There are several systems of romanization that is recognized in Japan, namely Kunrei, Nippon, and Hepburn [5]. In this study, the system used is the Hepburn.

Once converted into romaji, the text needs to undergo some processes, such as case folding, filtering, and whitespace removal. Case folding is the process to uniform case into lower case so that capital letters will be changed to lowercase. After uniformed into lowercase, the next process is filtering or stop words removal. In that process, things that are less relevant as punctuation will be eliminated. The last stage, namely whitespace removal will the abolished spaces and a new line from the text.

B. N-gram

N-gram is a character-based method [6]. The first step in the winnowing algorithm is the formation of n-grams. This method has input n', where this value will be the number of pieces or substring of text or words formed. This process took substring the characters are all n of sentence structure. This retrieval is done from beginning to end of document or text. If the n value is too big, the obtained words will be long. There will be a chance for important words to be uncaptured. However, if the n value is too small, the will be many chunks of words. There will be difficult to distinguish between words. With those reasons, the suitable value of n should be chosen which is the minimum value that can distinguish coincidental similarity [4]. N-gram is needed in winnowing algorithm as input to get certain words from the text [7].

C. Hashing

Hashing is the technique of transforming string the character string representing the value of the originals. The value of the hash or hash value the winnowing algorithm This represents the fingerprint of a text. The values obtained from the calculation of the ASCII value of each character. In this case, the characters in question are characters formed from N-gram.

Types of hashing used in this process are the algorithm of Rabin-Karp or called with rolling hash. This algorithm is used to solve problems related to string matching. Rolling hash is used to reduce the iteration when calculating hash value, making the process more effective [8]. A mathematical function of rolling hash is as in (1) below [9]:

$$H_{(C_2 \dots C_{K+1})} = \left(H_{(C_1 \dots C_K)} - C_1 * b^{(k-1)}\right) * b + H_{(C_{(k+1)})} \quad (1)$$

D. Windowing

Windowing is a process in the winnowing algorithm to save the results of the rolling hash that has been obtained in advance. Values hash grouped in the window of w, where this value needs to be specified in advance. The result of this windowing process is a series of values hashing all w that have been grouped.

E. Fingerprinting

The main process in winnowing algorithm is the selection of the fingerprint. This can be done by searching for the minimum hash value of every window. If in a window there are more than one hash with the minimum value, hash that was in the righter window will be chosen as fingerprint [4]. These values will then be assembled into a fingerprint. The value of the fingerprint This fingerprint will be used as a reference for calculating the similarity between the two texts.

F. Similarity Measurement

The last steps to determine the degree of similarity between the two texts is the fingerprint similarity. Input from this step is fingerprint of the two texts that will be compared. There are several methods that can be used to check the similarity between two texts or documents. The one used for this research is cosine similarity. The mathematic formula for cosine similarity is as shown in (2) below [10].

$$CS = \frac{|X \cap Y|}{|X|^{1/2}.|Y|^{1/2}} \times 100\% \quad (2)$$

X and Y are the fingerprint of the compared documents. The output value will be ranged from 0 to 1. It will be multiplied by 100% in order to get the similarity percentage.

III. Design of Automatic Essay Grading System with Winnowing Algorithm

On the research of this time, the system will be implemented using winnowing algorithm. The inputs of this system are the text which is the key answer to the problem, and the text is the answer from the students where both of them were short essay that only consist of 2-4 sentences. Both texts use the Japanese language. The text can be in the form of katakana, hiragana, or kanji. For the inputs to be well-processed, there should be some methods to prepare those input. The design in general as in Figure 1.

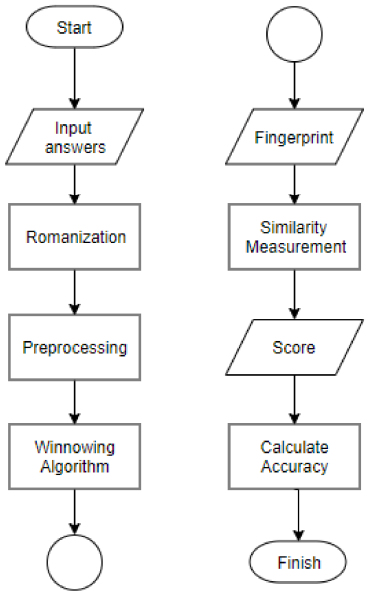


Fig. 1. Flowchart of SIMPLE-O with Winnowing algortihm

The first thing that the system do is initiating the essential inputs: student's answers and answer keys. However, those inputs are still in Japanese characters, such as hiragana, katakana, or kanji. There is a need for conversion from Japanese characters into roman letter because winnowing only utilizes ASCII in hashing process. This process called romanization. Moreover, with romanization process, kanji, hiragana, and katakana that could have the same pronunciation can be treated as the same.

After the romanization process was done to both documents, the next process is pre-processing. The purpose of this process is to prepare the data, making it easier for the next processes. It also helped with increasing the system accuracy. There are several things that will be done in this pre-processing stage: case folding, whitespace removal and filtering.

Case folding is a process to change the text into lower case. After that, the next process is whitespace removal. It removes space and new line from the text. The last in pre-processing stage is filtering. This process removes punctuation and special character, leaving only letter and number. The example of pre-processing is shown below.

- Japanese text : 彼は大学生です。
- Romanization : Kare wa daigakuseidesu.
- Case folding : kare wa daigakuseidesu.
- Whitespace removal: karewadaigakuseidesu.
- Filtering : karewadaigakuseidesu

After pre-processing, each input will be processed with winnowing algorithm. The output of this process is fingerprint. The fingerprint from the two texts will be compared to find the similarities using the similarity measurement method. The similarity percentage will be used for calculating the system accuracy by comparing it with the score from human rater. The formula used for calculating the accuracy is shown in (3) with a is the score obtained from system and b is the score from human rater. This accuracy will be used to find out the best winnowing parameter combination. The parameters are n in n-gram process, p in the hashing process, and w in windowing.

$$accuracy = 100 - abs(a - b) \quad (3)$$

This accuracy will be used to find out the best winnowing parameter combination. The parameters are n in n-gram process, p in the hashing process, and w in windowing.

IV. Results and Analysis

After the system was completely developed with the algorithms in the design process, it still needed some configuration to find the best running configuration to achieve the best result. The result is concluded as the best when the system could produce result that was close or similar to the result of human rater. So that in the testing process, it used answers that had 100% of score from human rater. As written in the purpose of this research, the system with winnowing algorithm parameters will be tested using varied values to find the effects on the obtained scores and system accuracy. The value of n and w will be varied from 2 to 7, while p will be varied to 2, 3, and 5. While one parameter is tested, the other parameters will be set into the value of 2.

A. Analysis on Variation of n Value

The value of n is varied from n=2 to n=7. From the graph shown in Figure 2, there is a decrease in accuracy along with the increase of n value.

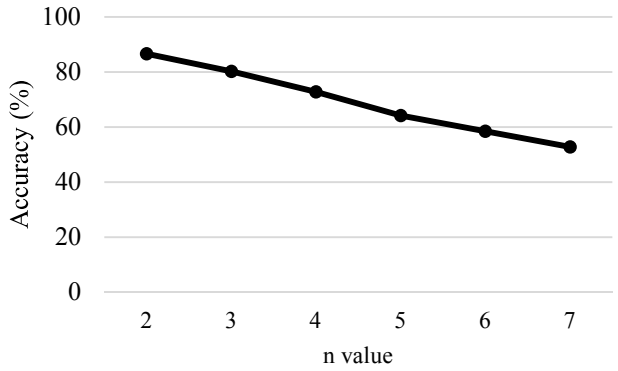


Fig. 2. Graph of effect of n variation on total score accuracy

TABLE I. Effect of n-Variation on Total Score Accuracy

n Variation	Accuracy (%)
2	86.86
3	80.49
4	73.20
5	64.58
6	58.80
7	53.18

According to Table 1, there is a decrease in accuracy along with the increase of n value or the chunk of words created. The highest accuracy from the system is 86.86% with n=2. The smaller the value of n, the shorter the words chunk, so more words chunk can be obtained. It increases the chance of finding the similarity between the two texts. The bigger value of n will make longer words, but fewer chunk.

B. Analysis on Variation of p Value

The value of p has variation p=2, p=3, and p=5. From the graph shown in Figure 3, there is a decrease in accuracy along with the increase of p value, but the accuracy is still above 85%.

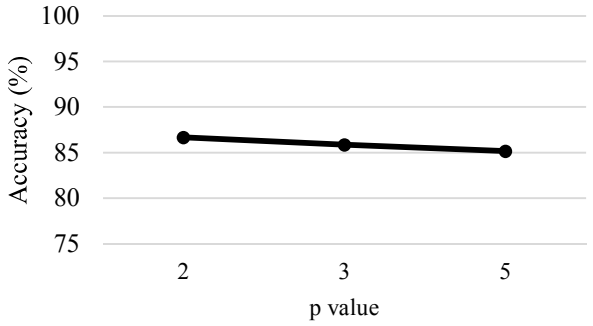


Fig. 3. Graph of effects of p-variation on total score accuracy

TABLE II. EFFECTS OF VARIATION ON TOTAL SCORE ACCURACY

p Variation	Accuracy (%)
2	86.86
3	85.86
5	85.41

According to Table 2, there is decrease in accuracy along with the increase of p value. This p value is a prime number used as multiplier in hashing process. The highest accuracy is 86.86% with p=2. The differences among the accuracy of the variation are not too significant. Different p value will affect the hash value obtained from the hashing process. As the result, there can be differences in fingerprint chosen from window in fingerprinting process.

C. Analysis on Variation of w Value

The value of w is varied from w=2 to w=7. From the graph shown in Figure 4, there is a decrease in accuracy along with the increase of w value, but the accuracy is still above 80%.

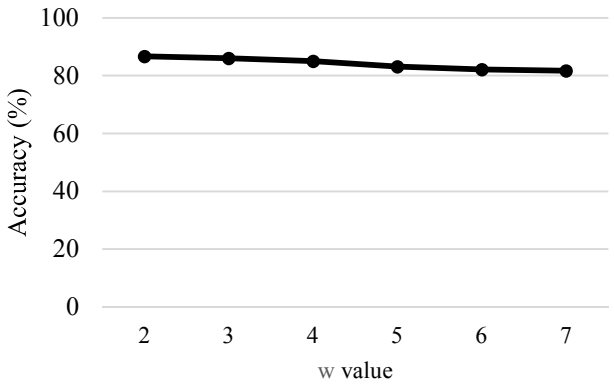


Fig. 4. Graph of effect of w-variation on total score accuracy

TABLE III. EFFECT OF W-VARIATION ON TOTAL SCORE ACCURACY

w Variation	Accuracy (%)
2	86.86
3	86.18
4	85.25
5	83.42
6	82.41
7	81.89

TABLE IV.

According to Table 3, there is decrease in accuracy along with the increase of w value. This w value is the window created in windowing process. The highest accuracy is 86.86% with w=2. The smaller the value of w, the shorter the window's length, so more windows can be created. It increases the chance of finding the similarity between the two texts. The bigger value of w will make longer window's length, but fewer in number.

D. Analysis on Variation Best Combination

In the experiments in each parameter, it can be seen but the parameter's variation can affect the system's accuracy. In the n variation's experiments, unlike p and w, there are variations with the accuracy below 80%. From this result, it can be said that the n parameter is better with small number.

According to the experiment, the most accurate parameters combination is n=2, p=2, and w=2. With this combination, the average accuracy is 86.86%.

In the previous research, the best parameters combination is n=2, p=3, and w=3. However, in this research, that combination gives accuracy that is lower than 86.86% from n=2, p=2, and w=2 combination

V. CONCLUSIONS

According to the experiments and analysis conducted, there are several conclusions that could be obtained. The smaller value of n in n-gram will result in higher student's score. The smaller

value of n also allows more fingerprint's similarity to be detected. The value of p in hashing affect the hashing process itself and the selected fingerprint from windows. The smaller value of p will result in higher student's score. The smaller value of w in windowing will result in higher student's score. The smaller value of w makes more windows can be created, also allows more fingerprints to be selected. The parameter combination with the highest average total score accuracy is n=2, p=2, and w=2 with 86.86%. With these results, this research group will continue the research as further work with that would be implemented into real application for real Japanese language exam in Japanese literature major study and improve the system with several parameter tunings and another methodology for improving its accuracy with input from the results of the exam. Design better user interface also become a top list to do so that it could ease general or non-technical people to use this application.

ACKNOWLEDGMENT

Our thanks to Universitas Indonesia and Ministry of Research, Technology, and Higher Education for supporting and funding this research under the grant of PITTA 2018 (Publikasi Internasional Terindeks untuk Tugas Akhir Mahasiswa) under contract number 2375/UN2.R3.1/HKP.05.00/2018.

REFERENCES

[1] A. A. P. Ratna, B. Budiardjo, and D. Hartanto, "Simple: Sistim Penilai Esei Otomatis Untuk Menilai Ujian Dalam Bahasa Indonesia," MAKARA of Technology Series, vol. 11, no. 1, 2007.

[2] B. Gipp dan N. Meuschke, "Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence", dalam Prosiding Simposium ACM ke-11 Tentang Rekayasa Dokumen (DocEng '11), Mountain, View, CA, USA, Sep. 2011

[3] K. Tung, N. Hung and L. Hauh, "A Comparison of Algorithms used to measure the Similarity between two documents", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET, vol. 4, no. 4, pp. 1117-1121, 2015.

[4] S. Schleimer, D. S. Wilkerson, and A. Aiken, "Winnowing," Proceedings of the 2003 ACM SIGMOD international conference on on Management of data - SIGMOD 03, 2003.

[5] W. Hadamitzky, Romanization systems. [Online]. Available: https://www.hadamitzky.de/english/lp_romanization_sys.htm#07. [Accessed: April 15, 2018].

[6] W. H. Gomaa and A. A. Fahmy, "A Survey of Text Similarity Approaches," International Journal of Computer Applications, vol. 68, no. 13, pp. 13–18, 2013.

[7] A. T. Wibowo, K. W. Sudarmadi, and A. M. Barmawi, "Comparison between fingerprint and winnowing algorithm to detect plagiarism fraud on Bahasa Indonesia documents," 2013 International Conference of Information and Communication Technology (ICoICT), 2013.

[8] R. Sutoyo, I. Ramadhani, A. D. Ardiatma, S. C. Bavana, H. L. H. S. Warnars, A. Trisetyarso, B. S. Abbas, and W. Suparta, "Detecting documents plagiarism using winnowing algorithm and k-gram method," 2017 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom), 2017.

[9] D. Purwitasari, I. W. S. Priantara, P. Y. Kusmawan, U. Y. Laili dan D. O. Siahaan, "The Use of Hartigan Index for Initializing K-Means++ in Detecting Similar Texts of Clustered Documents as a Plagiarism Indicator," dalam Asian Journal of Information Technology 10 (8-12), pp. 341-347, 2011.

[10] V. Thada dan D. Jaglan, "Comparison of Jaccard, Dice, Cosine Similarity Coefficient To Find Best Fitness Value for Web Retrieved Documents Using Genetic Algorithm", International Journal of Innovations in Engineering and Technology (IJIET), vol. 2, no. 4, pp. 202-205

Implementation Of Neural Network Backpropagation Using Audio Feature Extraction For Classification Of Gamelan Notes

Fahri Firdausillah
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
fahri@dsn.dinus.ac.id

Dwiky Gilang Mahendra
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
gilang.mahendra888@gmail.com

Junta Zeniarja
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
junta@dsn.dinus.ac.id

Ardytha Luthfiarta
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
ardytha.luthfiarta@dsn.dinus.ac.id

Heru Agus Santoso
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
heru.agus.santoso@dsn.dinus.ac.id

Adhitya Nugraha
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
adhitya@dsn.dinus.ac.id

Erwin Yudi Hidayat
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
erwin@dsn.dinus.ac.id

Abdul Syukur
Department of Computer Science
Dian Nuswantoro University
Semarang, Indonesia
abah.syukur01@dsn.dinus.ac.id

***Abstract*—Gamelan consists of several musical instruments including kendang, saron, bonang, panerus, kempul, gender, gambang, kethuk, flute, siter, clempung, slenthem, demung, japan, kempyang, peking, and gong. Many people do not know the name of each gamelan instrument and how it sounds. In an effort to increase the popularity and introduction of gamelan as a traditional musical instrument to the community, this research proposes an analysis based on audio classification of gamelan musical instruments. Classification is done by data mining technique using Backpropagation Neural Network (BPNN) method. Audio or sound data from gamelan recordings is conducted by preprocessing using Zero Crossing Rate and Short Time Energy before processed to Backpropagation Neural Network. This research only classifies 4 gamelan musical instruments namely gong, kenong, saron, and gambang. The experimental result shows accuracy at 82.5 %.**

Keywords—Classification, Audio Processing, Neural Network, Backpropagation

I. INTRODUCTION

Gamelan is a traditional musical instrument of Indonesia in the region of Java. Its surroundings, made of bronze, is a mixture of tin and copper. Gamelan presented in Javanese arts with a *karawitan* that portray a certain rhythm tone in harmony with the gamelan as a vocal singing accompaniment instrument by *sinden* (singer) [1]. Among the people, the popularity of gamelan is declining [2]. This happens because of the way to play gamelan is not flexible and the purchase price and maintenance of gamelan instruments are very expensive. While the number of gamelan masters is not that much, the popularity of the gamelan may decreasing [3].

Gamelan consists of several musical instruments including kendang, saron, bonang, panerus, kempul, gender, gambang, kethuk, flute, siter, clempung, slenthem, demung, japan, kempyang, peking, and gong [1]. Many people do not know the name of each gamelan instrument and how it sounds. Gamelan is closely related to hearing and feeling. To understand music or sound, a person must be well trained in his or her hearing and feelings in playing the tones. If the hearing is not trained to hear the sound of a tone, the gamelan music player is difficult to determine the tones he or she is hearing. Similarly, the feelings of gamelan music players. If his feelings are not yet integrated with the existing tones played, chances are, the player will play a tone with an incompatible tempo. Along with the development of this very fast era of gamelan as a culture and art of traditional, start to be forgotten.

There are researches about gamelan audio classification using data mining methods, such Naïve Bayes. Primary data is taken directly by recording the sound of gamelan music instrument. The audio feature extraction used in the research are Zero Crossing Rate and Short Time Energy. Based on the results of research methods Naïve Bayes used for the classification of sound gamelan music instrument has a good performance is evidenced with an accuracy of 92.5 % [2].

In an effort to increase the popularity and introduction of gamelan as a traditional musical instrument to the community this research proposes to make an analysis of audio classification of musical instruments gamelan. Each sound generated from each gamelan instrument has different characteristics that vary by sound frequency [4]. Classification is done by data mining technique using Backpropagation Neural Network (BPNN), as Backpropagation outperforms the results of Naive Bayes [5][6]. BPNN is one of the reliable algorithms in classification with various types of data [7].The

process of BPNN mimics the workings of the human brain that able to solve a problem by way of learning. BPNN is also able to adapt to learn from the input data provided, thus can map the relationship between input and output [8]. In addition, the ability of BPNN is well known in predicting output based on inputs that have been trained previously. This research uses BPNN as a method of gamelan audio classification and feature extraction of audio use Zero Crossing Rate and Short Time Energy. Classification is based on gamelan audio data that has different frequencies.

II. Related Study

There are several studies that serve as the study review in this study. Some related researches on gamelan and the use of neural network are described as follows.

In [2], the author discusses the use of data mining methods naïve bayes for gamelan sound classification. Primary data is taken directly by recording the gamelan music instrument. Gamelan instruments consist of kendang, saron, bonang, panerus, kempul, gender, gambang, kethuk, flute, siter, clempung, slenthem, demung, japan, kempyang, peking, and gong. According to the authors, many people do not know the name of each instrument of gamelan and how his voice. The sound of gamelan has different frequency in every gamelan instrument. Based on the results of research methods naïve bayes used for the classification of sound gamelan music instrument has a good performance with an accuracy of 92.5 %.

In [9], the author uses a radial basis function network for the identification of Javanese gamelan instruments. The method used for radial base network function training is the extended Kalman filter. Gamelan sound data are obtained from the recording of the three instruments of the saron group. There is no standard frequency in the ladder system of gamelan, as in modern musical instruments. This research was identifying Javanese gamelan using radial base function network. Radial basis function network is a multilayer feed-forward network whose training is hybrid. The results showed that the identification of Javanese gamelan instruments using radial basis function network trained using extended Kalman filters resulted in high accuracy of over 90 %. Thus the radial basis function network with training using extended Kalman filters can be used for the identification of Javanese gamelan instruments.

In [10], a system for voice recognition that focuses on numbers by applying MFCC and SOM Neural Network training is developed. Primary data is as input for training and testing steps. The system designed in this paper is a voice recognition application using the Mel method cepstral coefficients method to extract features and self-organizing maps to recognize sound. This application can recognize a voice that focuses on numbers. Based on the results of research the accuracy obtained by the proposed method is Mel frequency cepstral coefficients for extracting features and self-organizing maps to recognize notes of 82 percent.

Discussion of the classification of human voices is proposed by [11]. The algorithm used is Levenberg-Marquadt and Backpropagation Neural Network. Data is taken directly in the form of human voice recordings. The current speech recognition technology has improved especially in terms of speech processing. Speech processing is a way to extract the desired information from a sound signal. This study discusses the male and female voice classification system. Extracting the characteristics of the sound signals of each frame in the time zone and the frequency region is helpful to simplify and speed up the calculations. The features for voice or audio are Short Time Energy, Zero Crossing Rate, Spectral Centroid and others. The results of the system test show that the classification of human voice by using artificial neural network backpropagation and Levenberg-Marquadt algorithm for weight matrix changes, very good and fast because of the complexity of calculation that is not too high. Using 40 voices sample database with 5 test data, the result shows a classification that has been identified by a similarity value> = 0.5 as male and <0.5 as female. Tests using artificial neural networks resulted in average success rate in sound classification is 91%.

From the related study above, this research purposed a model to classify the sound of gamelan instruments using Backpropagation Neural Network and measuring its accuracy. To perform the audio classification of gamelan, Matlab is used. The data is the recording of the sound of gamelan instruments. The classification method is the Backpropagation Neural Network using audio extraction feature Short Time Energy and Zero Crossing Rate.

III. Experimental Result

A. Feature Extraction

The preprocessing stage is the processing of data from voice data into numerical values in order to be processed by BPNN method. The feature extraction stage consists as follows:

1. Short Time Energy(STE): marks the violence of the voice at a short time whos taken from

$$\frac{1}{N}\sum_{i=1}^{N} X(n)^2$$

- N: Number of samples / length of frequency
- X (n): The value of the signal from the sample

2. Zero Crossing Rate (ZC): denotes a sequential sample on a digital signal has a different sign, the size of a signal's noise in the domain feature whos taken from

$$ZC = \frac{\sum_{n=1}^{N} |sgn\ x(x) \text{-} sgn\ x(n-1)|}{2N}$$

- sgn x (n): the value of x (n), is 1 if x (n) is positive, -1 if x (n) is negative
- N: Number of Samples

Table 1 shows the feature extraction result of the gambang instrument. The table shows the file name, STE (Short Time Energy), ZC (Zero Crossing Rate) and Label.

Table 1: Dataset Extraction Features of Gambang

file	STE	ZC	label
gmb1 1.wav	0.0062	0.0293	gambang
gmb1 2.wav	0.052746	0.002318	gambang
gmb1 3.wav	0.052558	0.002319	Gambang
…	….	….	….
gmb9 5.wav	0.037425	0.010657	gambang

Table 2 shows the feature extraction result of the gong instrument. The table shows the file name, STE (Short Time Energy), ZC (Zero Crossing Rate) and Label.

Table 2: Dataset Extraction Features of Gong

file	STE	ZC	label
gong sedang 1.wav	0.005763	0.077479	gong
gong sedang 2.wav	0.005264	0.085212	gong
gong sedang 3.wav	0.005279	0.07921	gong
gong sedang 4.wav	0.005232	0.090106	gong
gong sedang 5.wav	0.005502	0.087265	gong
….	….	….	….
gs5 5.wav	0.00911	0.063908	gong

Table 3 shows the feature extraction result of the Kenong instrument. The table shows the file name, STE (Short Time Energy), ZC (Zero Crossing Rate) and Label.

Table 3: Dataset Extraction Features Data Kenong

File	STE	ZC	label
knp1 1.wav	0.028715	0.042177	kenong
knp1 2.wav	0.029188	0.041141	kenong
knp1 3.wav	0.028351	0.05063	kenong
knp1 4.wav	0.026956	0.044046	kenong
knp1 5.wav	0.027862	0.046402	kenong
…	….	….	….
….	….	….	….
kns6 5.wav	0.021367	0.057533	kenong

Table 4 shows the feature extraction result of the saron instrument. The table shows the file name, STE (Short Time Energy), ZC (Zero Crossing Rate) and Label.

Table 4: Dataset Extraction Features of Saron

file	STE	ZC	Label
sk1 1.wav	0.109531	0.0777	Saron
sk1 2.wav	0.109128	0.038031	Saron
sk1 3.wav	0.109064	0.04372	Saron
sk1 4.wav	0.107416	0.051157	Saron
sk1 5.wav	0.109092	0.06176	Saron
…	….	….	….
….	….	….	….
ss5 5.wav	0.041064	0.074181	Saron

B. Backpropagation Neural Network

1) BPNN Training

Research conducted in backpropagation network training using neural network parameters in Table 5. The amount of parameters depicted in the table are used due to the most optimal value regarding to an accuracy percentage and computing power, in such lower value will result lower accuracy but higher require more computing power without any significant improvement in accuracy.

Table 5: Backpropagation Neural Network Training Parameters

Parameter	**Value**
Epoch	10000
Learning Rate	0.01
Target Error	0.00000001
Hidden Layer	100 Neuron

Based on network parameters in Table 6 training is done with the process that can be seen in Fig 1.

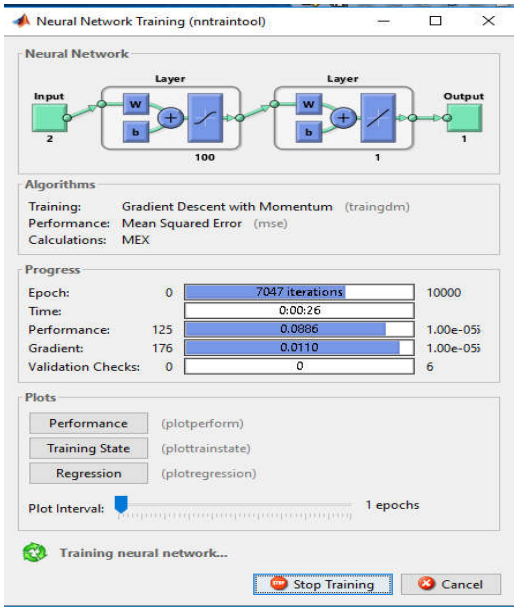


Fig 1. Backpropagation Neural Network Training

2) Implementation

This chapter describes the implementation of BPNN in the GUI using Matlab. Here are the steps of the classification that have been applied to the GUI.

a) Interface of Application

When the source code is executed, the initial view is shown in Figure 2. In the program view there are several components such as file formats and feature extraction results on the file.

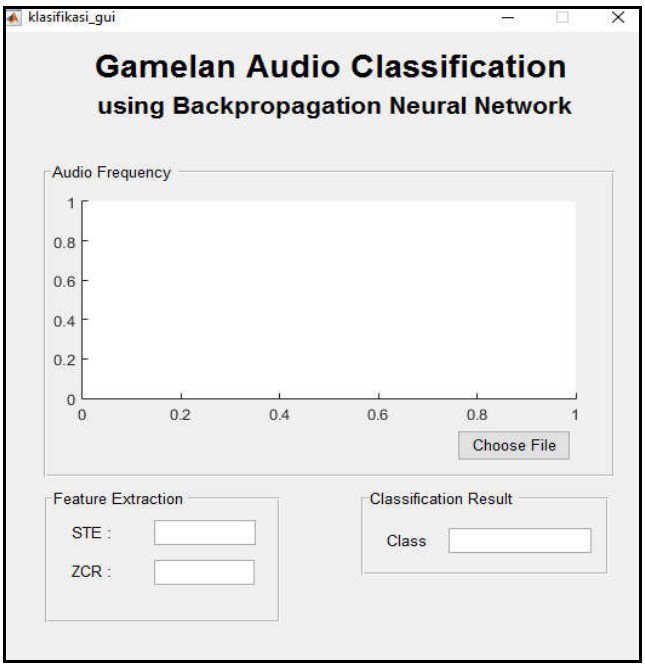


Fig. 2. Interface of Application

b) Audio File Selection

To be able to classify the initial stage is to select the file by pressing the file select button. Then the system will display a file selection dialog that is only in the filter file existence .wav

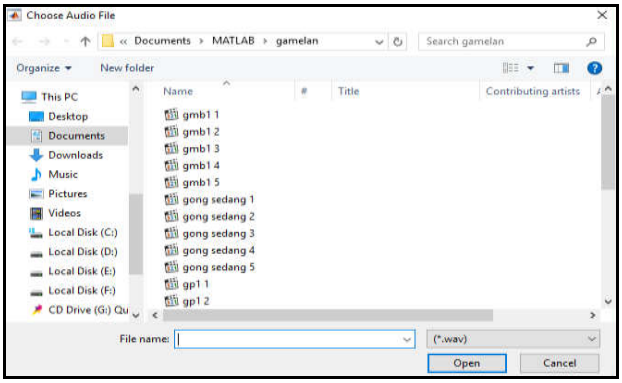


Fig.3. Select Audio File

c) Classification Results

After selecting the file, the system will automatically display all results starting from the feature extraction and the result of the classification of the selected audio file.

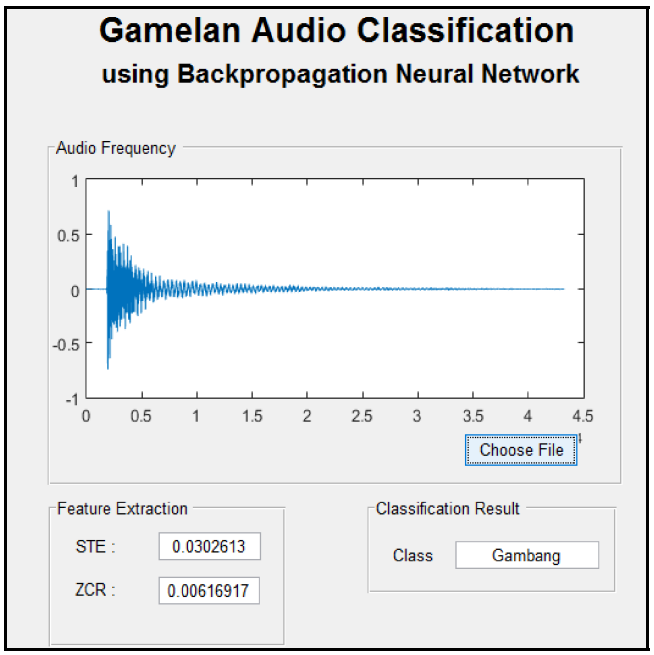


Fig. 4. Classification Result of Gambang

3) Evaluation

Once the Backpropagation Neural Network method has been successfully applied to the classification, tests are performed to determine the performance of the proposed method. The evaluation is done 40 times by using cross validation where training data is checked towards Backpropagation Neural Network model which has been made. After, the result of the test value is compared to available dataset.

Based on the test table results, confusion matrix is arranged to calculate the accuracy of the proposed classification.

Table 7. Evaluation Model using Confusion Matrix

		Class			
		Gambang	Gong	Kenong	Saron
Predicted Class	Gambang	9	0	1	0
	Gong	0	9	0	1
	Kenong	1	0	9	2
	Saron	0	1	1	6

$$Accuracy = \frac{True_{gambang}+True_{gong}+True_{kenong}+True_{saron}}{gambang+gong+kenong+saron} (1)$$

$$Accuracy = \frac{9+9+9+6}{10+10+10+10} = \frac{33}{40}$$

$$Accuracy = 82.5\%$$

Based on the experimental results that have been done in the initial process to produce the model. The model is then evaluated using confusion matrix to calculate the accuracy value. As shown in above calculation, the accuracy value of the evaluation reached 82.5%.

IV. Conclusion And Future Work

The proposed data mining method of Backpropagation Neural Network has been successfully applied for gamelan sound classification. The application of Backpropagation Neural Network method used for the classification in this study has an accuracy of 82.5 %.

Further research can apply and perform additional features used for classification so as to add the data calculation for the classification process to improve its performance. More in-depth research and other variations of methods for classification such as c45, Decision Tree, SVM, as well as optimization methods such as Genetic Algorithms and Particle Swarm Optimization are also another option.

References

[1] J. Lindsay, Javanese Gamelan, Oxford : Oxford University Press, 1St Edition edition, 1992.

[2] I. NUGROHO, "Analisa Dan Klasifikasi Suara Gamelan Menggunakan Algoritma Naïve Bayes," 2017.

[3] Y. T. C. Pramudi, F. Budiman and S. , "Desain Virtual Gamelan Jawa Sebagai Media Pembelajaran," in Seminar Nasional Aplikasi Teknologi Informasi 2010 (SNATI 2010), Yogyakarta, 2010.

[4] A. Rizky, "Identifikasi Instrumen Gamelan Jawa Menggunakan Jaringan Fungsi Basis Radial Dengan Metode Pelatihan Extended Kalman Filter," in Prosiding Seminar Nasional Matematika, 2014.

[5] M. Simfukwe, "Comparing Naive Bayes Method and Artificial Neural Network for Semen Quality Categorization," in International Journal of Innovative Science, Engineering & Technology, 2015.

[6] A. . A. S. "Heart Disease Prediction System using Multilayered Feed Forward Neural Network and Back Propagation Neural Network," in International Journal of Computer Applications, 2017.

[7] David, "Penerapan Algoritma Levenberg-Marquadt dan Backpropagation Neural Network Untuk Klasifikasi Suara Manusia," 2013.

[8] A. Hamid, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in IEEE International Conference on, 2013.

[9] A. Riski, "Identifikasi Instrumen Gamelan Jawa Menggunakan Jaringan Fungsi Basis Radial Dengan Metode Pelatihan Extended Kalman Filter," in Prosiding Seminar Nasional Matematika, 2014.

[10] M. Puspitasari, "Pengenalan Suara Menggunakan Mel Frequency Cepstral Coefficients Dan Self Organizing Maps," 2014.

[11] David, "Penerapan Algoritma Levenberg-Marquadt dan Backpropagation Neural Network Untuk Klasifikasi Suara Manusia," 2017.

Optimization of Evaluation and Monitoring Activities on Student Industrial Practices through Web-based Information System

Syaad Patmanthara
Departement of Engineering
State University of Malang
Malang, Indonesia
syaad.ft@um.ac.id

Wahyu Nur Hidayat
Departement of Engineering
State University of Malang
Malang, Indonesia
wahyu.nur.ft@um.ac.id

Khoirudin Asfani
Departement of Engineering
State University of Malang
Malang, Indonesia
khoirudin.asfani.ft@um.ac.id

M. Zainal Arifin
Departement of Engineering
State University of Malang
Malang, Indonesia
arifin.mzainal@gmail.com

Tri Atmadji Sutikno
Departement of Engineering
State University of Malang
Malang, Indonesia
tri.atmadji.ft@um.ac.id

Utomo Pujianto
Departement of Engineering
State University of Malang
Malang, Indonesia
utomo.pujianto.ft@um.ac.id

***Abstract*— This article aims on the user experience of web-based information system for optimizing evaluation and monitoring activities on student industrial practices. Web technology was chosen to be used in building the information system because of its reliability and accessibility. In this research, the information system has been built and tested for its user experience. Subjects of the product testing involved several students and experts. The students are about 50 samples and 2 expert assessments. Questionnaire was used for collecting data. The data consists of qualitative and quantitative. The result shows that the user experience is quite good, but the product need to be developed again based on the advice of the respondents.**

Keywords—information system, web, education, industrial practice

I. INTRODUCTION

Vocational education is required to develop social employment, maintenance, acceleration, and improve the quality of certain labors in order to increase community productivity [1] and it has a big role in planning and creating professional and productive human resources. The vocational education has different characteristics from general education, because it is organized to prepare ready to work graduates [2]. In the current era, the human resources (HR) demand who are skilled in their competence is absolutely necessary. To meet these conditions, the human resources quality of human resources must be improved.

In recent years, the system of Vocational Education and Training (VET) in some countries have been experiencing a renewal in order to fulfill society's demands [3][4]. This demand concerns how to design a qualified training program to work. The demand is caused by changes in the characteristics of the world of work and the qualifications of labors needed by the industry at this time [5]

One of effective vocational learning to produce human resources with the competence required by the world of work is industrial practice program. This program is a form of human capital investment and has its own uniqueness in terms of its education and training model. industrial practice program not only about how to learn skills in an academic setting but also learn skills in the workplace environment. Therefore, an information system that can accommodate the implementation of monitoring and evaluation of industrial practice needs to be developed as an effort to improve the effectiveness of such activities [6].

The information systems to monitor and evaluate industrial practice activities will be developed using web programming languages. The selection of the web as the basis of information systems due to web technology that is currently growing rapidly and can be accessed easily and conveniently through a smartphone (mobile device). As the growing educational needs, the implementation of IT Fusion is also inevitable.

II. IT FUSION IN LEARNING

The modernization of education is something that is happening in the midst of today's global era. Three things that need to be considered in change and modernizing education [7], i.e. (1) how; (2) what; and (3) when and where people learn. The role of information technology (IT) in education is very clear. The presence of IT with all its various levels has facilitated the change.

Information technology has three main functions used in education. First, the technology as tool for students and teachers to assist in learning. Second, technology as science where IT is part of the discipline that should be mastered by students and teachers. Finally, technology as material and aids for learning that interpreted as learning materials and also a tool to master a computer-assisted competence.

Web-based information system is one form of IT utilization in education, namely information systems to monitor and evaluate the activities of students' industrial practice. The goal is to increase the quality, efficiency, and effectiveness of monitoring and evaluating students' industrial practices and this is in line the statement on goals of IT utilization in education [8]. The utilization of IT in learning not only can increase the satisfaction and motivation

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

of its users to be able to achieve the expected competencies [9], lead the achievement of authentic assessment becomes possible. Authentic assessment is a format whereby students are assigned to apply their knowledge and skills in a tangible way [10].

The information system also must meet several criteria in order to meet the objectives of IT utilization in education [11]. These criteria consisting of system quality, quality of information, usability, user response, impact on users, and impact on relevant organizations/institutions. Therefore, testing of web-based information systems in this article will be analyzed based on these criteria.

III. Research Methods

This study uses research and development methods with a quantitative-qualitative approach. Initially, product development in this study used the waterfall method. But, to meet the needs in education, then the development model conceptually and procedurally used in this study adapts to Educational Research and Development (R&D), referring to research and development methods by Borg and Gall [12].

The research stage starts from collecting data of user requirements to producing products that have been tested and disseminated to the users. More detailed, there are ten steps that must be done, among others: (1) information collecting; (2) planning/design; (3) product development; (4) field-testing; (5) product revision; (6) main field-testing; (7) final product revision and development; and (8) implementation.

A. *Information collecting*

Data collection system needs information is done by distributing questionnaires to some lecturers and students who have done previous industrial practice.

The questionnaire contains statements regarding the shape of the information system and its functionality. Respondents are also given the opportunity to provide feedback in the form of suggestions in developing the information system. In general, information about user requirements is shown as in Fig. 1.

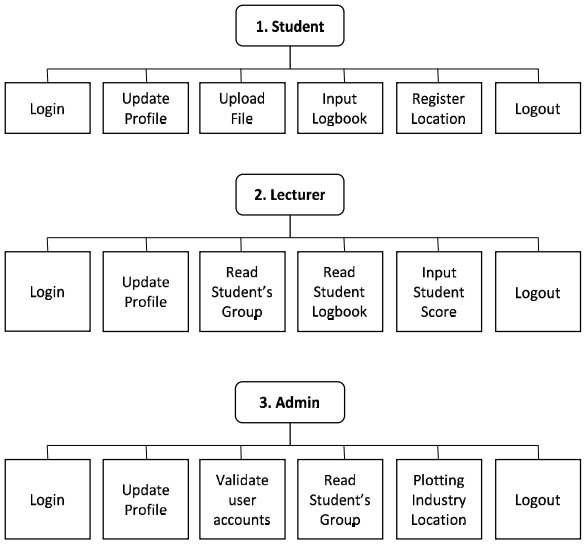


Fig. 1. Web's menus needs for system users

B. *Planning/design*

The development of web-based system is done by considering several aspects, namely the kind of users, functionality, and reusability. System users are divided into 3, i.e. web administrator, lecturer, and student. Based on it, the system user requirements are shown as in TABLE I.

TABLE I. Users Requirements

No	Type of User	Requirements
1	Administrator	1. Validate user accounts 2. Plot industry location 3. Read students group recapitulation 4. Read students distribution and statistic 5. Update profile
2	Lecturer	1. Read student's logbook 2. Enter student's score 3. Read/write message (chat) 4. Update profile
3	Student	1. Fill daily logbook 2. Register the location of industrial practice 3. Upload acceptance file (pdf/image) 4. Read students group recapitulation 5. Read/write message (chat) 6. Read evaluation score 7. Update profile

The research and development of this system is estimated to run for 3 months. Details of the product development time for a month, initial test conducted for 1 week, system revision phase 2 to 3 weeks, then final trial and its revision for 1 month. The design of the system development time estimation is made taking into consideration the timing of industrial practice activities that will soon be running, therefore the system must be completed in the near future.

C. *Product development*

Some software tools are used in developing this web-based information system. Web is programmed using Notepad++ and Sublime Text Editor. The database uses MySQL and Adobe Photoshop is used as image processing. Web development begins with building a database for information systems. The database is built by considering several parameters, such as user groups, data types to be used, to design Entity Relationship Diagram (ERD).

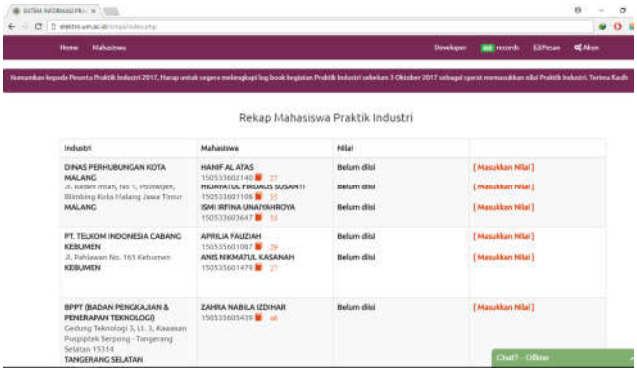


Fig. 2. Lecturer's students recapitulation page

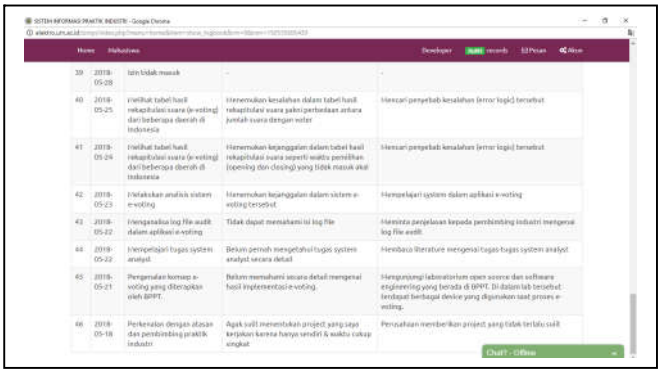


Fig. 3. Lecturer's student's logbook page

Once the database has been established and built, the next step is to build a web interface. Based on the design, each user created a web page as needed. After each user's web page has been completed and is ready for use, the initial web-based information system is ready for its functionality and usability testing.

D. *Field-testing*

Field-testing was conducted by involving 40 samples (38 students and 2 lecturer) who have completed industry practice. The test is done using the black box testing method. The method was chosen by considering the subjects of field-testing who have a role as users (non-developers). In addition, the focus of this field-testing is on product usability according to black box testing technique [13]. This technique is used to find out if the product is functioning properly. Representation of black box testing technique is shown as in Fig. 4.



Fig. 4. Black Box Testing Representation.

The black box testing stage begins with inviting respondents to use the web system directly. After using the web system, respondents were given the opportunity to fill out the questionnaire provided. Respondents filled out the questionnaire based on their real experience during the testing phase

The results of field-testing data consist of quantitative and qualitative data. Quantitative data is used as a measure of the level of system validity, while qualitative data is used as consideration in the form of suggestions. Based on the results of the field-testing, 90% score was obtained which means the system is eligible to use with a minor revision.

E. *Product revision*

Product revisions are based on field-testing results. The field-testing results show some things that need to be fixed, such as difficulties in performing features "forgot password", errors in updating data, and improvements in terms of its interface. Some of the menus need to be reorganized to make it easier for users to use the web-based information system.

F. *Main field-testing*

The main field-testing is done after the product has been revised before. This testing stage involved 75 students, 5 lecturers, and 2 experts as a consideration of the feasibility of web-based information systems. The result, both in terms of students and lecturers, almost all functionality goes well. A problem that arises is when students send files with "pdf" and "jpg" extension with naming files containing symbols, then the file will automatically go to the server but not detected.

G. *Final product revision and development*

Product development has been completed. Some improvements have been made. Expert advice in the form of web display improvements to be more responsive has also been implemented. Product development has been completed. Some improvements have been made. Expert advice in the form of web display improvements to be more responsive so the web system can be conveniently accessible and easy to use via smartphone. Furthermore, the web is ready to upload to the hosting server and ready for implementation in learning.

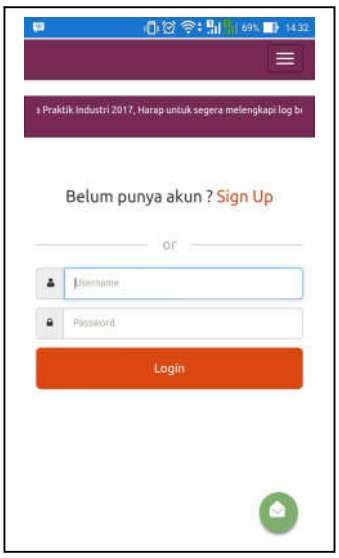


Fig. 5. Login page via smartphone

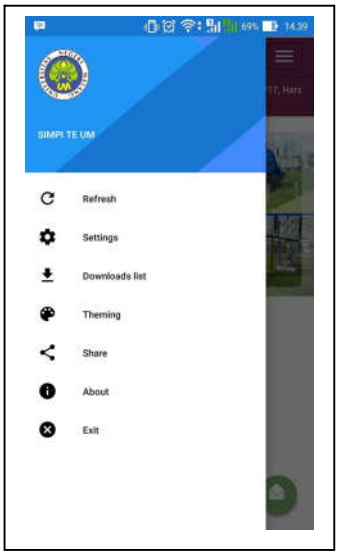


Fig. 6. Administrator's menu appearance on web page via smartphone

H. Implementation

Implementation of the product is done within approximately 2 months during the activity of industrial practice. When the implementation is in progress, the students and lecturers are given a questionnaire containing the user satisfaction. User satisfaction data adapts to the literature on web usability [14] which is shown in the aspects of efficiency, effectiveness, learnability, memorability, and error rate.

IV. Results and Discussions

A web-based information system for monitoring and evaluating student's industrial practice has been created. Implementation of monitoring and evaluation on students who are following industrial practice activities has been running for approximately 2 months.

User satisfaction data that has been given to a number of students and lecturers has been captured and analyzed. As a result, users both students and lecturers expressed satisfaction with the existence of information systems monitoring and evaluation of web-based industry practice with a feasibility average score of 83% (feasible based on TABLE II, with minor revisions). Recapitulation of user satisfaction questionnaire filling results presented as in TABLE III.

TABLE II. Feasible Criteria Percentage

No	Feasibility Level	Decision
1	81 - 100	Very Feasible with/without minor revisions
2	61 - 80	Feasible with minor revisions
3	41 - 60	Enough Feasible with revisions
4	21 – 40	Less Feasible with major revisions
5	0 - 20	Not Feasible

TABLE III. User Satisfaction Score

No	Aspect	Score (%)
1	Efficiency	90
2	Effectiveness	80
3	Learnability	80
4	Memorability	70
5	Error Rate	95
Average Score		83

Based on the recapitulation of user satisfaction data after using the web-based information system that has been made, the highest score is on the aspect of the error rate, then efficiency, followed by effectiveness and learnability, and the lowest score on memorability. Aspects of error rate is the highest score, it indicates that almost all the functionality that exist on the web such as login, update data, forgot password, entry values, logout, and so on goes well. There is only one malfunction in the system, that is, when the user wants to upload proof of acceptance of industry practice, all files with its size is over 2 MB will fail the process. Thus, it is necessary to check and re-examine the setting by increasing the amount of the maximum file's size that can be uploaded.

Efficiency becomes the second aspect that has the highest value of user satisfaction data. Based on the respondents' feedback (qualitative data), the users (students) get the ease of filling in their daily journals, consulting with their lecturers, and uploading the necessary files to support industrial practice activities into the system. Lecturers, as the students' mentor and assessor, also feel easier to monitor the work or daily agenda that Students do every day during their industrial practice. Student group's recaps can be clearly recorded. Communication becomes easier and integrated because it becomes one system. The assessment process is also easier and faster because when the score entered, all the assessment process will be automatically calculated in the system.

Next, aspects of effectiveness and learning ability become third order with the same score. This information system is considered effective because it helps users to be able to perform their tasks exactly in accordance with the expected goals. Students can upload their job report without having to bother to print and send the report to lecturer's locker. Students simply save their job report in a Microsoft Word's file, then upload it to the system so it can be accessed and assessed by the lecturer. Likewise, with lecturers, no need to bother visiting students to industrial sites. Lecturers simply send / leave messages to communicate and read as well as assess the file of each student's job report guidance appropriately. Usually, lecturers often lose reports that have been done by students or lecturers have difficulties to check the student work report because it is impossible to always bring the student work report. With the information system, all work becomes integrated, appropriate, and everything becomes as expected.

Learnability aspect also needs to be considered in developing an information system. Learning ability is an aspect of the ease of new users that able to interact and achieve optimal performance. This information system has a good learnability score. This is evidenced by the users'

response in learning the various menus and functionality provided by the web. Layout and naming menus, as well as icon shapes used in web content have been commonly used as they are in others. Other than that, most users (the majority of students also called Z generation) are able to quickly learn web functionality.

Finally, the lowest score is on the aspect of memorability. Although it's easy to learn and use, but the user's ability to reuse the web-based information system is not very good. Users still need guidebooks or wonder in advance when using or accessing the web. Generally, the users are getting used to the web after using it for the third and fourth time. However, this happens on average because they are afraid of making a mistake when they want to use the features available in the web-based information system.

V. CONCLUSIONS

In general, this research and development product produces a web-based information system that can optimize the performance of lecturers in monitoring and evaluating students who are doing industrial practice. Web systems that have been built are easy to use and easy to access (accessible) through a variety of commonly owned devices. Analysis of user satisfaction data states that the product developed is feasible and can be implemented.

The result, in terms of error rate, efficiency, effectiveness, learning ability, to memorability have a good score. However, as a follow up to this research and development, some suggestions from users need to be considered. Some features also need to be added, such as uploading massive data via Microsoft Excel's file. Furthermore, the product is expected to be developed in the form of mobile applications, so users do not have to bother to open the browser and type in the web address (direct access to the system).

VI. REFERENCES

[1] L. Clarke and C. Winch, Vocational Education International Approach, Development and System. New York: Routledge, 2007, pp. 62.

[2] A. Sonhadji, Manajemen Pendidikan dan Pelatihan. Malang: Universitas Negeri Malang, 2012.

[3] T. Deissinger, R. Heine, and M. Ott, "The Dominance of Apprenticeships in the German VET System and Its Implications for Europeanisation: A Comparative View in the Context of the EQF and the European LLL Strategy," Journal of Vocational Education and Training, vol. 63, no.3, 2011, pp. 397-416.

[4] I. Juul and C. H. Jorgensen, "Challenges for the dual system and occupational self-governance in Denmark," Journal of Vocational Education and Training, no. 63, vol. 3, 2011, pp. 289-303.

[5] E. Tome, "Employability, Skills, and Training in Portugal (1988-2000): Evidance from Official Data," Journal of European Industrial Training, vol. 31, no.5, 2007, pp. 336-357.

[6] S. Patmanthara and W. N. Hidayat, "The Effectiveness of Learning Management System (LMS) on Computer Assisted Learning Course for Informatics Engineering Education Students," Advanced Science Letters, vol. 24, no. 4, 2018, pp. 2642-2645.

[7] M. Resnick, "Rethinking Learning in the Digital Age," in The Global Information Technology Report: Readiness for the Networked World, G. Kirkman, Eds. Oxford: Oxford University Press, 2002, pp. 32-37.

[8] M. Alavi and R. B. Gallupe, "Using Information Technology in Learning: Case Studies in Business and Management Education Programs," Academy of Management Learning and Education 2003.

[9] K. Asfani, H. Suswanto, and A. P. Wibawa, "Influential factors of students' competence," World Transactions on Engineering and Technology Education, vol. 14, no. 3, 2016, pp. 416-420.

[10] J. Mueller, "The Authentic Assessment Toolbox: Enhancing Student Learning through Online Faculty Development," Journal of Online Learning and Teaching, vol. 1, no. 1, 2005, pp. 1-7.

[11] W. H. DeLone and E. R. McLean, "The DeLone and McLean Model of Information Systems Success: A Ten-year Update," Journal of Management Information Systems, vol. 19, no. 4, 2003.

[12] W. R. Borg and M. D. Gall, Educational Research: An Introduction 5th. ed. New York: Longman, 1989.

[13] M. Kumar, S. K. Singh, and R. K. Dwivedi, "A Comparative Study of Black Box Testing and White Box Testing Techniques," International Journal of Advance Research in Computer Science and Management Studies, vol. 3, issue 10, 2015, pp. 32-44.

[14] W. Quesenbery, "The Five Dimension of Usability," in Content and Complexity: information Design in Technical Communication, Michael J. Albers and Mary Beth Mazur, Eds. New York: Routledge, 2013.

Implementation of Weighted Naive Bayes Algorithm for Major Determination in Indonesian High School

Erwin Yudi Hidayat, Aulia Sabiq Taufiqurrahman, Ardytha Luthfiarta, Junta Zeniarja, Heru Agus Santoso, Adhitya Nugraha
Faculty of Computer Science
Universitas Dian Nuswantoro
Semarang, Indonesia
erwin@dsn.dinus.ac.id, auliasabiq@ymail.com, ardytha.luthfiarta@dsn.dinus.ac.id, junta@dsn.dinus.ac.id, heru.agus.santoso@dsn.dinus.ac.id , adhitya@dsn.dinus.ac.id

Rismiyati
Department of Computer Science
Universitas Diponegoro
Semarang, Indonesia
rismi@if.undip.ac.id

***Abstract*—Major selection for 9th grade students aim to focus the student on building the skill and passion that they owned. It is also aimed to build the students' potential, passion and talent. By selecting the major that students passionate about, it is expected that they can improve the score of academic students. The process of major selection today has many drawbacks. One of them is that the process only considering the preference of students without looking at the students' academic background. This resulted in possible problem for students in the future, for example the students will not be able to pass the grades. In addition, the major selection will also affect the possible course when the students decide to continue their study to university level. This study proposed major selection by applying the techniques of classification in data mining. Weighted Naive Bayes Algorithm is implemented to determine high school major of students in junior high school of 33 Semarang. Experimental result shows 80% in accuracy. The system prototyping is developed for the visualization that can help student to determine major in senior high school.**

Keywords—Data Mining, Classification, Weighted Naïve Bayes Algorithm, Major Determination

I. INTRODUCTION

Education is an important aspect in every country's development. In Indonesia, the government has implemented 12-year compulsory education for the students. Twelve-year education consists of 6-year primary school, 3-year junior high school and 3-year senior high school. From primary to Junior high school, every student will take the same subject. However, during upon entering Senior high school, every student will need to choose major.

The purpose of major selection for 9th grade students is to enable the student to focus on building their skills and specific interest. By focusing on their interest and skill, it is expected that they can improve their academic scores. In addition, there are efforts from the government to monitor and regulate the education system as best as possible, so that parents have no difficulty in guiding their children [1].

Major selection conducted today has many disadvantages. The first disadvantages is that this process only consider students' preference without considering the academic background of the students. This will cause some problem in the future, for example, students' academic grades are not optimal, selection programs when students proceed to university level are limited due to inappropriate secondary school majors, and others [2]

The selection of majors in students' aims to facilitate students achieve national education goals. In this context, principals and counseling teachers have an important role in providing services to students. This means that service to students requires special professional handling, in order to support the career success of students in the future. Errors in determining departmental services for students can have a negative impact on student learning processes and output [3].

Considering the importance of choosing majors for students, a method is needed that can help minimize the impact of mistakes when choosing a major. Data mining is a branch of science that studies methods to extract knowledge or patterns from data (discovery of knowledge in a database) [4]. The process of data classification, predictions, estimation, and determination of association rules in the data can be done using Data Mining. Data Mining is needed because of the large amount of data that can be used to produce useful information and knowledge [5].

Data mining technique can be used to obtain strategic information that can be used to determine the selection of majors by the students will be obtained. The information obtained through the analysis results can be used to find the new opportunities and finding the strategic plan in the process. However, the activities of classification carried out by the human still have limitations, especially on the human ability to accommodate the amount data that you want to be processed. In addition, the error due to inaccuracy made by the human may also occur. A possible solution to handle this problem is by applying data mining techniques that can be used for processing the data into strategic information resources. Data mining can help an organization that has the overflow data to provide more information to support decision making [6].

Many researchers typically using the Bayesian Network for data mining [7]. It is Naive Bayes, a very practical network model [8]. Bayesian theorem is the theorem referred to by

Naive Bayes. Although it is fairly simple, it has a loss, which is when the attributes are conditionally independence [9]. Determining the weight value for each individual attribute can reduce the main disadvantages of Naive Bayes. This is because the weight value will force the attribute to have different tasks in the classification process. As a result, Weighted Naive Bayes participates in alleviating conditional independence, so that Naive Bayes is more practical in real-world applications implementation [10].

II. LITERATURE REVIEW

Classification is a process to identify the class from one case based on attributes collection. Research which aims to improve classifier accuracy level is a very good research topic in the field of machine learning and data mining. One of the most effective methods of classification is Bayesian Network which was introduced by Pearl. Conditional probabilities and network system are the component of A Bayesian Network. The system is described in the form of acrylic directed graph node that corresponding with the variable domain, while direct dependency between variables is represented by arcs between nodes. Naive Bayes is a classification technique that is customarily used in data mining and machine learning. This method proved to be very effective to solve problems in a collection of classification data. However, in real world applications, violation is often occurred to the strong assumption that all independent attributes. Reducing the assumption of independent attributes is one of several methods to improve performance of Naive Bayes. On the other hand, the expected error may arise due to the violation of the independence assumption. Assigning the weights to each attribute is another way improve the Naive Bayes [11].

Bayesian Network is a method that often used in the data mining processes, one of them is classification. Naive Bayes is a model from Bayesian Network which is very simple. Naive Bayes refers to Bayesian theorem and simple model but has the independent attributes. As we know, the assumption made by the independent attributes of Naive Bayes harms the performance of the classification when it is violated in real life. Undercut the assumption of independent attributes on Naive Bayes and maintain the simplicity and efficiency attributes of the Naive Bayes, can improve its performance. Some well-known effective methods are such Naive Bayes Decision-Tree Hybrid, Tree Augmented Naive Bayes, and Selective Naive Bayes [12].

Data mining is a technique that can be used to explore valuable information. This information was originally hidden, or invisible in a very large data collection. This technique is able to find unique patterns, which were not known beforehand. The word mining itself has the meaning effort to get a bit valuable items from a large number of base materials. So that data mining has long roots of disciplines such as artificial intelligence, machine learning, statistics, and databases. Genetic algorithm, neural networks, association rules mining, classification, and clustering are some methods that are oftenly mentioned in the literature of data mining [13].

Bayesian Network is a graphical model based on graph theory and probability theory called Bayesian network. It forms a relationship model between the domain variables under study. Bayesian network provides an approach to reasoning directly by manipulating the probability. It is more important for machine learning as well as data mining, and has been implemented to a wide range of jobs, including a medical expert system, natural language understanding, manufacturing of errors' diagnosis, and user modeling [10]. Bayes theorem:

$$P(H|\mathbf{X}) = \frac{P(\mathbf{X}|H)P(H)}{P(\mathbf{X})} = P(\mathbf{X}|H) \times P(H)/P(\mathbf{X}) \quad (1)$$

Naïve Bayes is a simple Bayesian network. Naïve Bayes is generally used for problems in the classification process [10]. Naïve Bayes is easy to design, because it has a very simple system [9]. Naïve Bayes does not use complicated rules, decision trees or other explicit representation. On the other hand, to find the most likely of the possible classification, theory of probability under the branch of mathematics is can be used [9]. Naïve Bayes presents a solution that combines a prior probability and the conditional probabilities in a single formula. Thus, to calculate the probability of each classification may use this method. After that, select the greatest value in the classification process [9]. Naïve Bayes formula can be derived from Bayes Theorem:

$$P(C_i|X) = \frac{\mathrm{P(X|C_i)P(C_i)}}{\mathrm{P(X)}} \quad (2)$$

Bayesian network is a method in data mining that is frequently used by researchers [9]. Bayesian network model that is very practical is Naive Bayes [10]. Naïve Bayes refers to Bayesian Theorem and has simple system, but they have a major disadvantage if the attribute conditionally independence [11].

$$P(c_i|X) = \underset{c_i \in C}{\arg} P(c_i) \prod_{j=1}^{n} P(x_j|c_i)^{w_j} \quad (3)$$

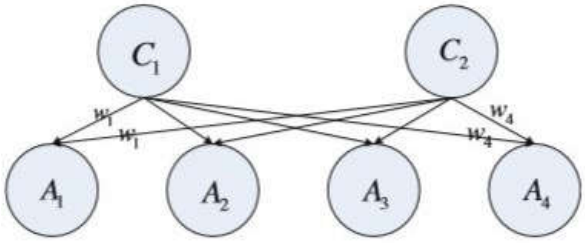


Fig. 1. The framework of simple Weighted Naïve Bayes

As mentioned beforehand, determining the weight values for each single attribute is one of the main ways to reduce the main drawback of the Naïve Bayes. Because the value of the weight pushes the attributes to take on different characters in the classification, so the Weighted Naive Bayes can support to ease the assumption of conditional independence and make

Naive Bayes more practical for applications that applied in human life [10].

Fig. 1 shows the formula and the framework for simple Weighted Naïve Bayes. Here is the example of manual calculation in implementing Weighted Naive Bayes to determine student major using some training dataset that has been prepared. This calculation will produce the result whether the student will be categorized as IPA or IPS using Weighted Naive Bayes algorithm. These are the following steps of manual calculation for Weighted Naive Bayes Algorithm:

1. Prepare data training
2. Calculate the mean value of each attribute

$$\mu = \frac{x1 + x2 + \cdots + xn}{n} \quad (4)$$

3. Calculate the standard deviation of attribute in each class

$$s = \sqrt{\frac{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_1)^2}{n(n-1)}} \quad (5)$$

4. Calculate the weight of each attribute

$$w = \left| \frac{\mu(IPA) - \mu(IPS)}{\sigma(IPA) + \sigma(IPS)} \right| \quad (6)$$

5. Calculate using Weighted Naïve Bayes formula

III. Methodology

A. Data

This research utilize data that is retrieved from junior high school of 33 Semarang. The data that provided by the school is in xls format. The student data that comprise of student names, their choice of senior high school, majors, and the score of the student from a questioner. This is raw data that is used in this study. This raw data contains 1000 student's records, each of which has 13 attributes. The example of the raw data is shown in Table I.

TABLE I. Student Raw Data

No		NAME	DESTINATION	SEX	SCORE						INT.	CLASS
SEQ	TEST				BI	ENG	MATH	NS	SS	ICT		
1	305	AAN ZAKARIA	SMAN 15 SEMARANG	M	92	80	82,5	80	81,7	83,3	IPA	IPS
2	523	ABDUL MUIS	SMAN 11 SEMARANG	M	88	87	85	87,9	88,5	87	IPA	IPS
3	512	ACHMAD TAUFIQ	SMAN 11 SEMARANG	M	87,7	85	82,6	81,6	80	93	IPA	IPA
4	422	AENI VERAYANTI	SMAN 2 SEMARANG	F	93	93,2	83	83,4	87	88	IPS	IPA
5	446	BAROYAH USWATUN	SMAN 1 SEMARANG	F	91,5	87	85	84,3	91,8	85,5	IPA	IPA
6	42	DENI KUSUMA	SMAN 11 SEMARANG	M	92,5	80	85	85	85	80	IPS	IPA
7	486	DERINTAH MELATI										
8	517	DERY SUMARGO	SMAN 11 SEMARANG	M	79,2	76,7	73,3	75,8	76	85	IPS	IPA
9	171	DEWI ATIKA	SMAN 2 SEMARANG	F	84	83	87	82,49	75	85	IPA	IPA
10	290	DEWI SARTIKA	SMAN 15 SEMARANG	F	84,6	81,7	82,1	82	86,7	83,3	IPA	IPA
11	169	EGA RIZKY FAJAR	SMAN 15 SEMARANG	M		96,3	82	84,3	92,5			IPA
12	303	FALA AULIA	SMAN 1 SEMARANG	F	93	95	90	87	77	81	IPA	IPS
13	427	FAZHIA DAMAYANTI										
14	555	FAWAZZI MUHAMMAD	SMAN 11 SEMARANG	M	80,9	83,5	80,8	81	90	88,3	IPS	IPS
15	469	HIDAYATUS SOLEHAH	SMAN 11 SEMARANG	F	94	92,3	92,6	92,2	88	84	IPA	IPS
16	350	IIS KAROMAH	SMAN 15 SEMARANG	F	90	86	85	72,5	84	83	IPS	IPS
17	27	IKA WIDIASTUTI		F	86,6	81,5	82,5	87	77	84,6		
18	473	INDAH NOVITASARI	SMAN 2 SEMARANG	F	90,7	89,1	88	89,3	90	90	IPA	IPA
19	324	KARINA SAHARANI	SMAN 2 SEMARANG	F	91	97	82	88	80	92	IPA	IPA
20	245	HARIMAN SIREGAR	SMAN 15 SEMARANG	M	90	90	70	75	95	90	IPS	IPS
21	83	MUHAMMAD BRYAN	SMAN 11 SEMARANG	M							IPA	IPA
22	356	KINANTI JASMINE	SMAN 15 SEMARANG	F	85	92	80	78	93	80	IPA	IPS

B. Method

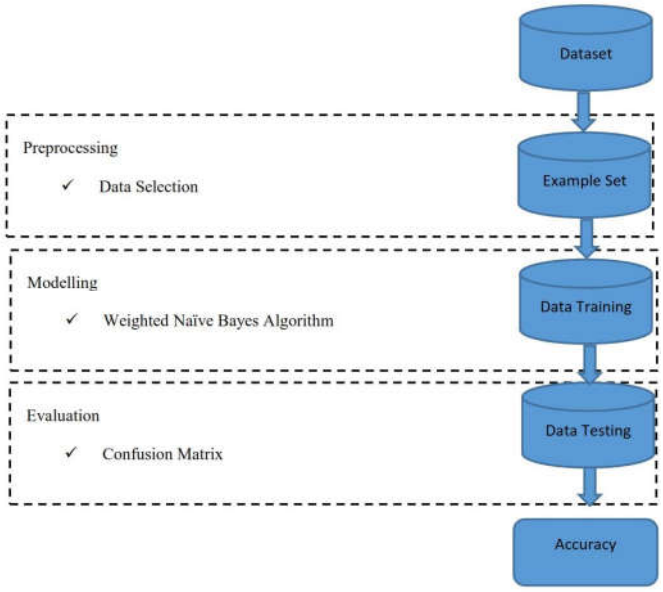


Fig. 2. Research Flow

The flow of the research conducted in this study is shown in Fig 2. The first step of the research is data collection followed by data preprocessing. Data preprocessing consist of data selection, data reduction, and data transformation. The next step is data split whereby data is divided into training set and testing set. Data training will be trained by using Weighted Naïve Bayes algorithm. The resulting model will be tested against testing data where the accuracy of the model will be measured.

IV. Result and Discussion

A. Preprocessing

The raw data cannot be directly used as the experimental data. It is because of the original student data do not have enough predictor variable attributes that become the requirements of supervised learning. It means the data preparation need to be performed in the very beginning of the experimental process before doing a testing. There are some steps taken to get the original student data become an example set used in this study. Those are data selection, data reduction, and data transformation.

Table II explained that data transformation is performed by changing some of data to make data mining process become easier after the dataset are selected. The data that will be transformed are Indonesian language or Bahasa Indonesia (BI), English (ENG), Mathematics (MATH), Natural Science (NS), and Social Science (SS). Based on curriculum in 2013, the score that obtained by student can be written in number score or letter score. For example, 90 is can be converted into A score. From these regulations, the final score that obtain by the student can be converted into five possibilities, those are A, B, C, D, and E. The value of D and E will cause the student cannot pass the exam. On the other hand, if students achieve A, B, or C the student certainly pass the exam.

Because values of variable are transformed into 5 categories, the data transformation which are 0-29,99 = E, 30-44,99 = D, 45-69,99 = C, 70-84,99 = B and >85 = A to get the nominal value from numeric value. These scores range based on junior high school of 33 Semarang regulation. Each school has different score range, so each school has different rules to transform the student score. In junior high school of 33 Semarang, the student who could not got A or B score, they could not pass the exam and should did remedial exam until they reached A or B score.

The result of data preprocessing is shown in Table II. From 1000 raw data, only 907 can be used for this research. Each data has 7 attributes as shown in Table II. The remaining data cannot be used because there are missing values in some of the variable.

B. Naïve Bayes Classifier

The preprocessed data is split into training data and testing data. Total training data used is 750 data, while the remaining 157 will be used as testing data. Fig. 3 shows the example of classification result using Weighted Naïve Bayes algorithm. Data 1 as a dataset for the process. The data testing is the score of Bahasa Indonesia, Mathematic, Computer, English, Science, and Social that has been transformed from nominal score. The result of given data testing based on classification model is Science.

Fig. 4 shows the detailed calculation process of Weighted Naïve Bayes algorithm of Fig. 3. The final value of Science is 0.13142 and Social is 0.04036. So, from these results, the classification result based on above calculation is Science.

TABLE II. EXAMPLE OF DATASET

BI	ENG	MATH	NS	SS	ICT	CLASS
A	B	B	B	B	B	IPS
A	A	A	A	A	A	IPS
A	A	B	B	B	A	IPA
A	A	B	B	A	A	IPA
A	A	A	B	A	A	IPA
A	B	A	A	A	B	IPA
A	B	B	B	B	B	IPS
B	B	B	B	B	A	IPA
B	B	A	B	B	A	IPA
B	B	B	B	A	B	IPA
A	A	B	B	A	A	IPA
A	A	A	A	B	B	IPS
A	A	B	A	B	B	IPS
B	B	B	B	A	A	IPS
A	A	A	A	A	B	IPS
A	A	A	B	B	B	IPS
A	B	B	A	B	B	IPA
A	A	A	A	A	A	IPA
A	A	B	A	B	A	IPA
B	B	B	B	B	B	IPA

The confusion matrix for 157 testing data is shown in Fig. 5. Based on the given confusion matrix, the accuracy of the system can be calculated as follow

Accuracy = (true Science + True social)/(total data)
=(72+54)/157 = 80.25%

Based on the above result, it is shown the best testing result shows the result of 80.25% accuracy. With the Social or Natural Science classification 90% accuracy and IPS or Social Science classification 70% accuracy.



Fig. 3. Result of Calculation Process



Fig. 4. Result of Calculating Weighted Naïve Bayes

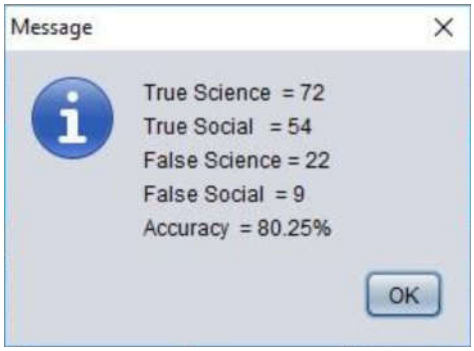


Fig. 5. Confusion Matrix

V. CONCLUSION

In this study, the implementation of Weighted Naive Bayes Algorithm for classification prediction to determine the major

in senior high school shows some result. The data used in this study is from SMA 33 Semarang. Total data used in the experiment is 907 data, which consist of 750 training data and 157 testing data. The best testing result shown the result of 80% accuracy. Thus it can be concluded that weighted Naïve Bayes Algorithm is the one method that can be applied to determine the major in senior high school that is used to help the junior high school student. Further research can be performed on the same problem by introducing another classifier to improve the accuracy result.

REFERENCES

[1] Premono Agung,"Kompetensi Keahlian Sekolah Menengah Kejuruan : Antara Kebijakan dan Realita",2010.

[2] Nurmansyah MH and Rahayu Y, "Sistem Pendukung Keputusan Penjurusan SMA Menggunakan Simple Additive Weighting", 2012.

[3] Fauzan A. Maharani. (2014, Juli) http://www.m- edukasi.web.id. [Online]. http://www.m- edukasi.web.id/2014/07/peningkatan-mutu-pendidikan- smasmk.html?m=1

[4] Witten Ian H., Eibe Frank, and Hall Mark A., Data Mining : Practical Machine Learning Tools and Techniques 3rd Edition. Elsevier, 2011.

[5] Obbie Kristanto, "Penerapan Algoritma Klasifikasi Data Mining ID3 untuk Menentukan Penjurusan Siswa SMAN 6 Semarang," 2014.

[6] D. Kiron, R. Shockley, N. Kruschwitz, G. Finch, and M. Haydock, Analytics: The Widening Divide, MIT Sloan Management Review, 53(2), 1-22, 2012.

[7] Pearl, J,: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Reasoning, Morgan Kaufmann Publishers, Los Altos, 1988.

[8] Langley, P., Iba, W., Thompson, K,: An analysis of Bayesian classifers, AAAI 90, 223–228, 1992.

[9] Xie, Z., Hsu, W., Liu, Z., Li Lee, M,: SNNB: a selective neighborhood based Naïve Bayes for lazy learning, In: Chen, M.-S., Yu, P.S., Liu, B. (eds.) PAKDD 2002, LNCS (LNAI), vol. 2336, pp. 104–114, Springer, Heidelberg, 2002.

[10] N. A. Zaidi, J. Cerquides, M. J. Carman, and G. I. Webb, "Alleviating Naive Bayes attribute independence assumption by attribute weighting," Journal of Machine Learning Research, vol. 14, pp. 1947–1988, 2013.

[11] Sona Taheri, John Yearwood, Musa Mammadov, Sattar Seifollahi, "Attribute Weighted Naïve Bayes Classifier Using a Local Optimization," 2013.

[12] Xuesong Yan, Wei Li, Qinghua Wu, Victor S. Sheng, "A Double Weighted Naive Bayes for Multi-label Classification," 2016.

[13] Turban Efraim, Aronson Jay E, and Liang, Decision Support Systems and Intelligent Systems, 7th ed.: Prentice Hall, Upper Saddle River, NJ, 2005.

Heart Beat Based Drowsiness Detection System for Driver

Prima Dewi Purnamasari, Aziz Zul Hazmi
Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, Depok, 16424, Indonesia
prima.dp@ui.ac.id

***Abstract*—Most traffic accidents are caused by negligence of the driver in managing rest time when driving; driver was sleepy and so could not control the vehicle. Therefore, it is necessary to have a device that can detect sleepiness and warn the driver beforehand so that the driver can avoid the accident. The development of this drowsiness detection system is using the heartbeat as a source of data to be retrieved using Photoplethysmography sensor. The system then classifies and determines the drowsiness level of the driver based on his/her heartbeat. This system uses Arduino Nano and Odroid XU4 as the processing unit and has LCD to display the output. The developed system has the success rate up to 96.52%. From this work, it has been proven that a person's sleep condition will affect the heart rate.**

Keywords—drowsiness detection; Photoplethysmography; heart beat.

I. INTRODUCTION

Since humans know how to ride, humans continue to develop their ability to drive from time to time. From using animals as vehicles, then wheels are found, until a machine for a vehicle is found, humans always develop their driving abilities by various purposes, such as shortening travel time from place to place to support work or human activities. But with the development of human vehicles, other problems arise such as accidents caused by the lack of safety of the vehicle or even accidents caused by the driver itself, especially about sleepiness problems when driving.

Based on data from Republic of Indonesia traffic police corps (Korlantas Polri), the number of traffic accidents in 2017 reached 102,057 cases where the cause of the accident is still dominated by tired and drowsiness. This shows that tired and drowsiness when driving in Indonesia is still very frequent. Of the number of accidents also make a huge loss, in terms of material and the victim itself. The total number of accidents in the last quarter reached 24,011 cases from all over Indonesia with death toll of 5,902 dead, 3,504 injured, and 27,938 people slightly injured with total loss reached Rp 51,130,883,831 [1].

Sleepiness and Tiredness is natural if experienced by someone especially for people who have high mobility such as inter-city bus driver or truck driver who usually they drive in a long distance and a long time. In article 90 of Law Number 22 of 2009 paragraph (3) it is mentioned that public transport drivers are required to rest half an hour after driving for four consecutive hours. In the same article, in paragraph (4), it is explained that in some cases the driver may be employed no later than 12 hours a day, including an hour-long rest period [2]. Although in some circumstances the rider has been anticipating to have a replacement driver if it is getting sleepy, but still many cases that occur due to the driver forced to keep driving even though his condition was sleepy. This can lead to accidents that will harm many parties. Therefore, it is necessary to detect sleepiness that can be used by the driver so that the device can warn the rider if the device has been getting signs that the driver was sleepy. To make the drowsiness device, the tool will be made using a heartbeat detection using a sensor that will detect the heartbeat and analyze whether the heart rate indicates that the driver is sleepy or not. This research aims to be implemented in real driving conditions; however, for safety reason, the experiment and testing presented in this paper was conducted from non-driving person.

This paper is structured as follows. Section II explain about theories that support this research include the heartbeat, and some hardware that are used in this developed system such as Odroid XU4, Arduino Nano and Photoplethysmography sensor. Section III discuss about methodology that we implemented in this research; the design of both hardware and software. Section IV discuss about experiment and its result. Section V concludes the paper.

II. LITERATURE REVIEW

A. Heart Beat

The pulse rate is the number of times the arteries (blood vessels clean) to expand and contract in one minute as a response to the heartbeat. The number of pulses equals the heart rate. This is because heart contraction causes an increase in blood pressure and pulse in the arteries. Measuring the pulse is synonymous with measuring the heart rate.

In a normal heart, the heart rate is the rate at which the sinoatrium node depolarizes as it is the source of cardiac depolarization. Heartbeat, like other vital signs such as blood pressure and respiratory rate, changes with age. In adults, a normal heartbeat is between 60 and 100 beats per minute (normocardic) in which the children are taller. An abnormal heartbeat is called bradycardia (<60 in adults) and higher than usual is tachycardia (> 100 in adults). The complication of this is when the atria and ventricles are out of sync and the "heart rate" should be determined as an atrium or ventricle (eg, the ventricular level in ventricular fibrillation is 300-600 bpm, whereas the atrial level can be normal (60-100) or faster (100-150)) [3].

Meanwhile, when someone enters the relaxation phase the heart rate will be reduced by about 8 beats per minute. This is obtained by determining the distance from the average heart rate to the minimum heart rate at bedtime on each subject tested by Miriam R. Waldeck (2003) [4]. This causes a decrease in blood pressure, which occurs in deep sleep. During rest, blood flows from the brain, relaxes the arteries and enlarges body tissues. A decrease in heart rate can also be caused by the parasympathetic nerves of the heart that begin to activate when a person starts to drowsy. At the heart, there are two autonomic nerves that function to regulate the cells that work under the consciousness of the person and the two nerves are also present in the heart because the heart works autonomously. The autonomic nerves are divided into two sympathetic and parasympathetic nerves. The sympathetic and parasympathetic nerves work opposite, the sympathetic nerves work to improve the work of heart cells such as when exercising or when the person is intense, while the parasympathetic nerves work to lower the work of the heart cells as they are at rest. When we rest the parasympathetic nerves will actively work to lower the work of heart cells so the body can rest.

B. Photoplethysmography

Photoplethysmography (PPG) is a simple optical technique used to detect volumetric changes in blood in the peripheral circulation. This is a low-cost and non-invasive method that makes measurements on the surface of the skin. This technique provides valuable information related to our cardiovascular system. Recent technological advances have aroused interest in this technique, which is widely used in clinical physiological measurements and monitoring [5].

The PPG waveform consists of the physiological waveform of the pulsatile ('AC') associated with the synchronous changes of the heart in the blood volume with each heartbeat and superimposed on a slowly varying 'DC' basis with the various low-frequency components attached with respiration, sympathetic activity of the nervous system and thermoregulation. Although the origin of PPG signal components is not fully understood, it is generally accepted that they can provide valuable information about the cardiovascular system [6]. The illustration of PPG mechanism is depicte in Fig 1.

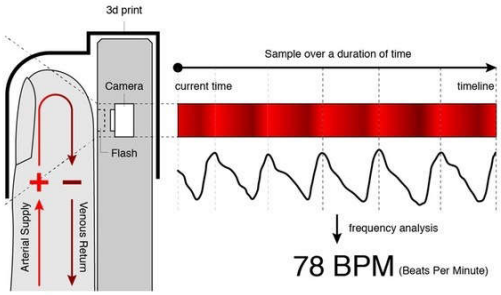


Fig. 1. **Example of PPG signals when detecting heart rate** [7]

In the PPG signal there is some information that will be processed until it can be a heartbeat calculation that is Peak, Through, and IBI (inter-beat-interval). The PPG signal is processed by detecting Peak from the PPG analog signal obtained from the sensor and calculating the interval between the beats. This intermittent interval called IBI is used to estimate a heart rate such as estimating the average heart rate of a few beats.

The main advantage of the PPG sensor is that the signal gets stronger in movement compared to other heartbeat sensors, the price of PPG sensors is also not too expensive when compared with other sensors, the treatment of these sensors is also easy so it is suitable for portable equipment and does not require big power.

To estimate how many heartbeats, the signal reading will be taken every two milliseconds so that the signal sent by the sensor can be processed optimally. If the number of signal readings has reached 250 times and read continuously then the new sought how many heartbeats with some conditions. If the signal is larger than the threshold and the signal readout value is greater than 3/5 IBI (3/5 IBI is used to avoid noise from the crosses signal), from this initial condition an IBI time benchmark will be used next. The second IBI value will be obtained when a second heart rate has been obtained. The second IBI value will be stored for use in the next calculation. The process will continue until there are 10 IBI data which will be averaged and used to get BPM by dividing it by 60000 ms to get the correct BPM [8].

The PPG used in this system is pulse sensor SEN-11574. This pulse sensor uses a light sensor from Avago (APDS-9008) that is used as a heartbeat reader. The sensor reads the heartbeat from the amount of light obtained from the LEDs next to it. Because when the heart pumps blood it will make blood cells will increase and make the light intensity will be reduced. After the light intensity decreases the resistance value of the light sensor will increase [9].

The basic working principle of this sensor starts with the LED mounted on the sensor will light up after VCC and GND get the voltage. Secondly after there is a finger covering the LED and the light sensor, the light sensor will read the change in the intensity of light caused by the pulsation of the heart. When the heart beats there will be blood collecting to make the LED light slightly dimmed. Finally, the signal obtained by the sensor will be forwarded to the signal cable for processing [10].

C. Odroid XU4

Odroid XU4 is one type of microcomputer board (single-board circuit). It has the size of a credit card and has a digital output input port like on a microcontroller board. Odroid can be used to run office programs, computer game programs, and as a media player to high resolution video. Excess Odroid than microcontroller board that has a port / connection for display in the form of TV or PC Monitor and USB connection for Keyboard and Mouse.

The goal of creating an Odroid is to create low-cost devices that will improve programming capabilities and hardware understanding at pre-university level. Due to its small size and affordable price, it can be quickly adopted by electronic enthusiasts for projects that require more than basic microcontrollers (such as Arduino devices).

Odroid is open hardware, except the main chip on the Odroid is the Samsung Exynos544 Cortex-A15 + A7 2 GHz Cortex (Octacore 64 bit) that can run many major components of the board - CPU, graphics, memory, USB controller, etc. Many projects are created using Odroid and are well documented so we can create and modify them ourselves. Odroid is designed for many operating systems such as Linux, Armbian, Linux Times, Android, Arch Linux, and Lakka [11].

D. Arduino Nano

Arduino Nano is a development board (development board) microcontroller based on ATmega328P chip with a very small shape. The function is no different than Arduino Uno. The main difference lies in the absence of DC power jack and the use of Mini-B USB connector. Called as a development board because this board does serve as a means to create a prototype of a microcontroller circuit. By using the development board, you will more easily assemble the microcontroller electronics circuit than if you start assembling ATMega328 from scratch on the breadboard [12].

For this program Arduino Nano is used as a slave of a system already created in Odroid. This is because of the limitations possessed by the Odroid in reading analog signals on pins owned by Odroid. The Odroid analog pin can only be used with C Language while the program was originally created in Python. The use of Arduino Nano also makes it easier to read heartbeat as it has more sources of documentation that can be used as a reference when writing this program. Arduino Nano is sorted because for this project it does not need many pins, so Arduino Nano alone is enough to meet the need of pins needed.

III. METHODOLOGY

A. Hardware Design

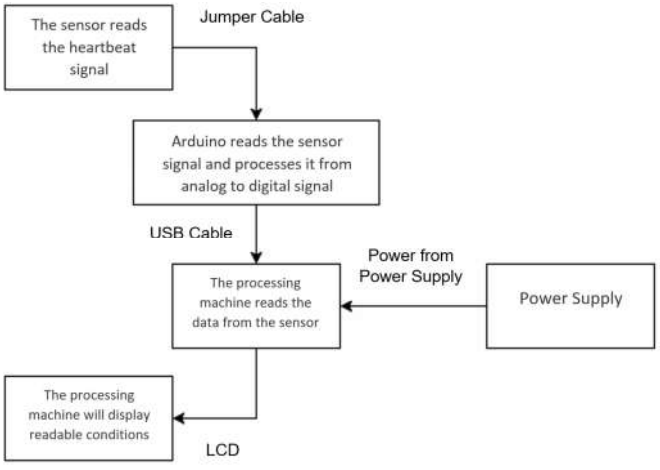


Fig. 2. Hardware Block Diagram

Heart rate sensors are used in a prototype of drowsiness detection system is useful for retrieving data from the driver's heart rate through the skin surface which then signals the data will be sent to a processing machine for processing. The signal generated by this sensor is an analog signal, so to facilitate the processing of this data, this analog signal must be converted first into a digital signal using Analog to Digital Converter that is in Arduino Nano.

First, the system will process the received PPG signal from the sensor to calculate how much BPM at the beginning of the calculation. The things that will be used for the calculation in this system is the value of IBI, after the IBI obtained then can be searched how many BPM owned by that person. then the system and sensors will perform the calibration by calculating the average heart rate in the first 5 seconds the user puts his finger to the sensor and buzzer will light up when the average heart rate in 5 seconds has been obtained. Furthermore, the screen will display the average heart rate which means that the calibration is successful. The system will run and will take the heart rate in real time. The sensor will process the heartbeat and compare it with the lower limit of the user's already established heart rate after the average is obtained. The lower limit (hb^-) is determined by reducing the mean heart rate by eight. If the heart rate is 8 beats below the average that has been obtained before, then the system will determine the object is in a state of sleepiness. If the object is in a drowsy condition ie when 8 beats below the average or the heart rate is not legible for more than 2 seconds, then an alert are displayed on the screen to warn that the user is in a drowsy state.

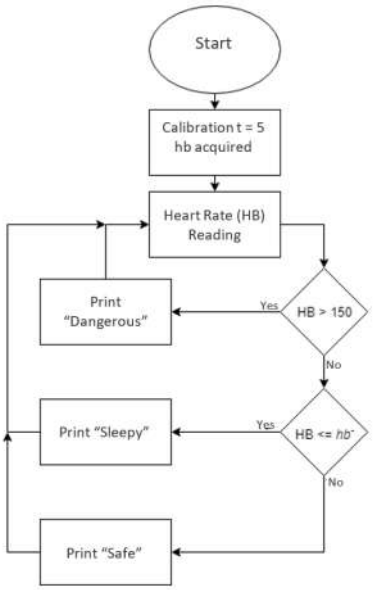


Fig. 3. Overall System Flowchart

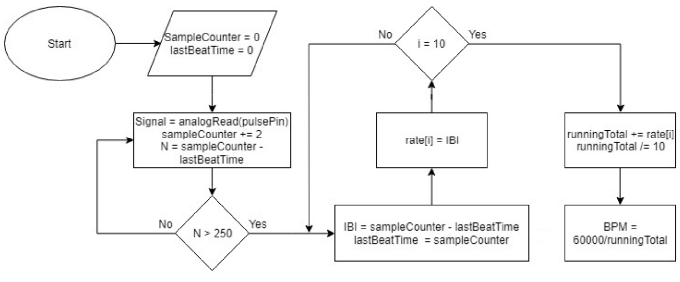


Fig. 4. Heart Beat Detection Flowchart

The flow chart in Fig. 3 is an overall flow chart for a drowsiness detection system. More detailed explanations such

as the calibration function, heartbeat detection, and detection process and analysis will be explained at the fractions of the next flow chart. The BPM detection function overall is shown by the flowchart in Fig. 4, where the PPG signal sent by the sensor and is processed and calculated until the BPM is obtained.

Before the system is used to detect drowsiness, first, the system should be calibrated by calculating the average of 10 data received from the sensor. The average value will be the threshold to determine the drowsiness condition of a person. Then Odroid XU4 will read the heartbeat input from the sensor. The heartbeat of the sensor using the PPG algorithm. Determination of the threshold of a normal heartbeat based on the average data (the calibration data) already obtained by reducing the average heart rate with the value eight, hb^-.

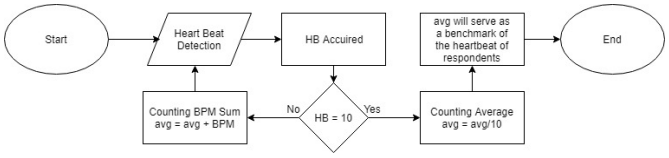


Fig. 5. Calibration Phase Flowchart

Fig. 6 is a flow chart of the heartbeat detection process. Where this process will involve the user directly after the user put his finger on the sensor. First, the system will read the signal data generated by the sensor from the previous program results. Furthermore, the system will process the data to get the calculation of how many beats per minute. The data pulse per minute will be compared with the calibration data already obtained previously. Heart rate obtained from the system will be processed and developed into a form of signal formed by Photoplethysmography so that the heartbeat readings will become easier.

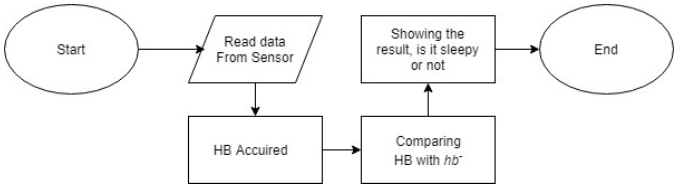


Fig. 6. Heart Beat Detection Phase Flowchart

IV. Experiment and Analysis

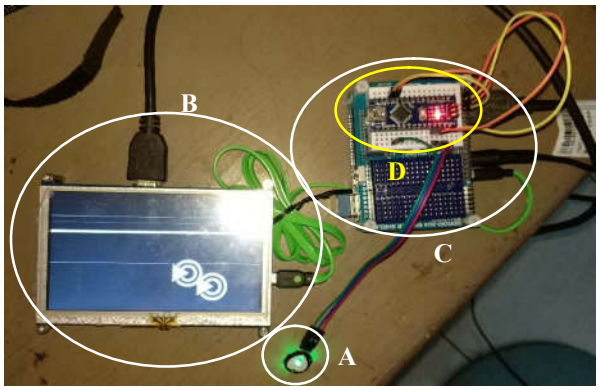


Fig. 7. Prototype of the System Hardware

Fig. 7 shows the prototype of the hardware for the heart beat based drowsiness detection system. Point A shows the sensor which will be attached to the respondent's finger wrapped with a velcro strap. At point B there is an LCD screen to display information about the respondent's heart rate and his/her condition. The LCD used is a 7 "LCD with 800 x 480 resolution that uses an HDMI cable for a graphical interface to the Odroid XU4 and micro USB as a power source. Points C is an Arduino Nano placed on top of the Odroid XU4 which is useful for reading data from the sensor and sending the data to the Odroid. The Arduino Nano is placed on a small breadboard for easy replacement. Point D is an Odroid XU4 which is the processing center for this drowsiness detection system.

This system was tested using three different types of experiments, that was designed to produce the best measuremet of this system. The first experiment is to compare this system with other trusted systems or devices that deliver true heart rate results, i.e. consumer MI Smartband that can display the heartbeat of the user and a medical device that OMRON Automatic Blood Pressure Monitor which can display the blood pressure can also display the heartbeat of the user. The second experiment was carried out in different respondents' conditions; i.e. in a fit (not sleepy), drowsy (sleepy), and sleep conditions. While the last experiment was to compare different lower limit of the heartbeat that is considered as sleepy condition.

A. System testing compared to benchmark device

The first experiment is to perform tests to compare the system to other machines that can measure heart rate: a smartband Xiaomi Mi Band 2, and a digital Sphygmomanometer OMRON Automatic Blood Pressure Monitor. The data retrieved by this experiment is used to determine the accuracy of the developed system by comparing it with a trusted/common tool to measure heart rate. The greater the accuracy of this system the better the results of the system already developed.

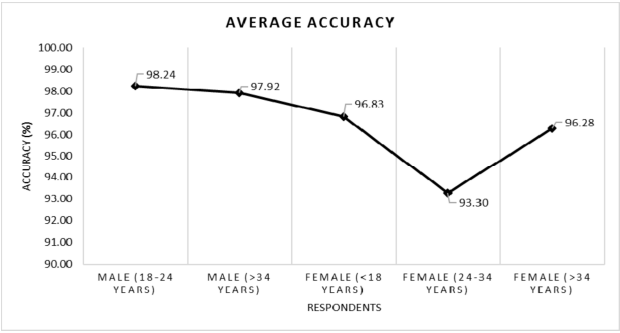


Fig. 8. Average Graph of Accuracy Among Respondents

Based on experiment result in Fig. 8, it is found that the developed system has a good accuracy because the difference between the system made with OMRON Automatic Blood Pressure Monitor tool is only 3.48%. This proves that the calculation of the heartbeat sensor is correct even though there is still some data that is quite different from the digital Sphygmomanometer. The cause of reading data on female respondents aged 24-34 years is lower than others can occur

because in the last experiment there was a slight shift in the finger of the respondent so as to make the reading by the data slightly disturbed.

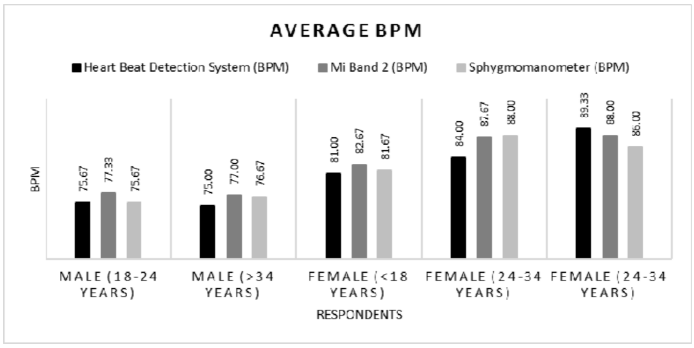


Fig. 9. Average BPM among Respondents Using Benchmark Devices

Whereas when viewed from BPM each respondent, this result depends on the health condition of each respondent. Rapid heartbeat can be caused by many things such as activity, hormones, gender, and age. From Fig. 9, it can be seen that the heartbeat of women is higher than the heartbeat of men, and age also affects the heartbeat. But this can happen also because of the diseases of each respondent, male respondents aged> 34 years have a history of high blood disease and have weight above 90 kg while women aged 24-34 years have a history of low blood disease and only have a weight around 40-50 kg, these two things can be one factor why their heart rate difference is quite high.

B. *System testing for different condition: not sleepy, drowsy and sleep*

The second test is conducted to see the difference in heartbeat for the three condition of the respondents: not sleepy, drowsy and sleeping. This experiment aims to prove whether a change in a person's awareness condition will affect the person's heart rate and how far one person's heart rate changes in a state of not sleepy, drowsy, and sleep.

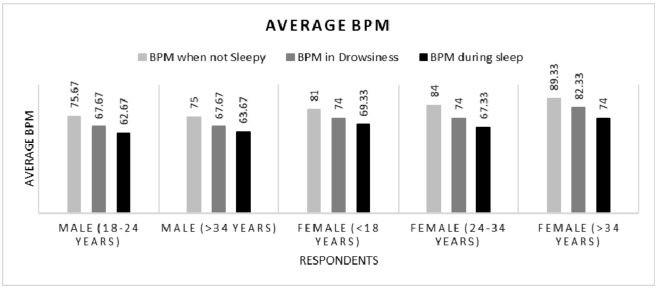


Fig. 10. Compariosn of Heartbeat in different condition: not sleepy (fully conscious), drowsy, and sleep

Based on the results in Fig. 10, it is found that the BPM of each respondent will decrease according to their conditions. Decreased calculations are obtained by reducing the average heart rate when fresh with an average heart rate when drowsy/sleep and then divided by the average heart rate when not sleepy, so it will be known how many decrease in heart rate that occurs from sleepy and sleep condition.

From the experiment, it was found that the decrease of each respondent was 7-11% from fresh to drowsy and 14-17% from fresh to falling asleep. This proves that when a person's body is in a state of relaxation then automatically their heart rate will also go down. The decrease that occurs is also quite in accordance with the reference where the human heartbeat will be reduced by about eight beats as it enters the relaxed phase [4]. More precisely in this study the heart rate of a person will drop from 7-10 beats from the beats as they are not sleepy. While in the sleep phase, the heart rate of a person will drop from 11-16 beats from the beat when not sleepy. This proves that when humans enter the relaxation phase the heart rate will decrease and the heart rate will also decrease further if it has entered a deeper sleep phase.

C. *Comarison of different treshold setting*

The third experiment is performed using a variation of the threshold to determine which lower limit would provide more suitable warning if drowsiness has been detected. This experiment was conducted using three types of lower boundaries there are -5, -8, and -10. This experiment is done by sending data down regularly as a simulation when someone is drowsy. The heart rate will start at 150 BPM and will continue to fall below the specified threshold data. After the heartbeat reaches the lower limit of the threshold it will count the time it takes to reach the limit of the heartbeat. The experiment started when the Odroid began receiving data submissions from Arduino Uno.

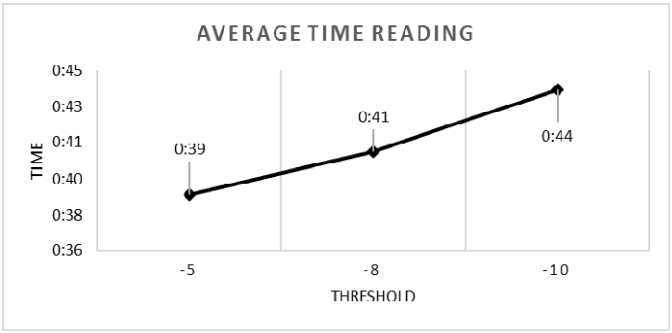


Fig. 11. Average Time Reading Graph

From Fig. 11, it can be seen that the results of the reading time obtained will continue to increase in accordance with the specified threshold; if the specified threshold is raised then the time will also increase. Of the three thresholds compared, the most appropriate result is the threshold -8 because the time reading is more suitable for the system compared to the other two thresholds. At the threshold -5, the system will read too early in the drowsy determination, so that it can make a person who is not yet sleepy will be detected as drowsy by the system. Whereas, the threshold -10 will cause the system to be late in detecting the sleepiness; such as the system wrongly recognize a person to be drowsy, although he is actually already in a sleep state.

V. CONCLUSION

From the experiments, it is proven that the developed system using PPG sensor is able to detect a person's sleep

condition, especially for driver. The ability of this heartbeat detector is good with a success rate of 96.52% of the total respondents, but there is still a problem in terms of user connections with the sensor. If the sensor is not moved from the fingers when reading the data then the data obtained will be quite good. So, it takes the binder on the finger when this sensor is in use. The lower limit of the heartbeat to determine someone drowsy or not more suitable is made at -8 to prevent premature detection or late detection with a detection-average detection time of 41 seconds.

The developed heartbeat based drowsiness detection system presented in paper already shows a good result. However, to improve the accuracy of the drowsiness detection, this system may be improved by using other physiological sensor, such as EEG or by combining with other method, such as image processing.

ACKNOWLEDGMENT

The author would like to thank the Ministry of Research and Higher Education of Indonesia for supporting this research under the PDUPT 2018 grant no 417/UN2.R3.1/HKP.05.00/2018.

REFERENCES

[1] Korlantas Polri, "Statistik Laka," *Korlantas Polri*, 2017. .
[2] M. A. Kusmagi, "Membedah Undang-Undang Lalu Lintas dan Angkutan Jalan Nomor 22 Tahun 2009," in *Selamat Berkendara di Jalan Raya*, Raih Asa Sukses, 2010, p. 15.
[3] American Heart Association, "All About Heart Rate (Pulse)," *heart.org*, 2015. .
[4] M. R. Waldeck and M. I. Lambert, "Heart rate during sleep: Implications for monitoring trainig status," *J. Sport. Sci. Med.*, vol. 2, no. 4, pp. 133–138, 2003.
[5] S. Cheriyedath. MSc, "Photoplethysmography (PPG)," no. February, pp. 1–22, 2010.
[6] G. Langereis, "Photoplethysmography (PPG) system," no. February, pp. 1–22, 2010.
[7] M. Elgendi, "On the Analysis of Fingertip Photoplethysmogram Signals," *Curr. Cardiol. Rev.*, vol. 8, no. 1, pp. 14–25, 2012.
[8] BioSec.Lab, "Photoplethysmograph (PPG) based Biometric Recognition," *Biometrics Security Lab*, 2017. .
[9] Osram, "SFH 7050 - Photoplethysmography Sensor Application Note," vol. 7050, pp. 1–11, 2014.
[10] Pulsesensor.com, "Open Hardware," *pulsesensor.com*, 2012. .
[11] Hardkernel.co, "Odroid-XU4," *Hardkernel.co*, 2013. .
[12] Circuits Today, "Arduino Nano V3.0 Pinout," *Circuits Today*, 2018. .

Economic Studies of the Wind Turbin-Diesel Hybrid Power Generation System (case study at: Queen of the South Beach Resort Hotel, Yogyakarta, Indonesia)

Husein Mubarok
Electrical Engineering Dept.
Universitas Islam Indonesia
Yogyakarta, Indonesia
mubsrok.husein@uii.ac.id

Tegar Krisna Yoga
Electrical Engineering Dept.
Universitas Islam Indonesia
Yogyakarta, Indonesia
13524031@students.uii.ac.id

Abstract—**The hotel operations require electrical energy and PLN becomes the main supply of electrical energy in Indonesia. But PLN is a power supplier company that can not avoid the disturbance or damage. Diesel generator is a generator that is commonly used as a backup generator at the hotel because it has advantages that are: a fast starting time and does not require a large area. However, the operational costs and the fuel of diesel generators are the greatest among other conventional types of generators. This research proposes wind turbine-diesel hybrid generator as an economical hotel generator replacing the use of diesel generator. An optimization process is required to realize that purpose. This study uses HOMER software to optimize the specification of the generator, especially the amount of wind turbine (WT), inverter and battery to obtain an economical hotel reserve power. The assessment's criteria used are the economic value of the plant : the Net Present Cost (NPC), Cost of Energy (COE) and Renewable Penetration are also calculated to determine the absorption of renewable energy sources by the hotel reserve power. The calculations are divided into two scenarios, scenario 1 and scenario 2. In scenario 1 the energy uses grid components and diesel generator while in scenario 2 energy uses grid components, diesel genertor, inverter, battery and wind turbine.The results showed that the most economical wind turbine-diesel hybrid system generation for the hotel in this study consisted of 1 unit of 200 kw diesel generator, 1 unit of 125 kw inverter, 7 units of 10 kw wind turbine and 1 unit of 50 kw battery. The economic value of wind turbine-diesel hybrid system is more economical than using diesel generator, where NPC system is reduced by Rp 3,470,852,394 or 21.75% compared to using only diesel generator with COE value which also decreased by Rp 570 or 33.17%. The wind turbine-diesel hybrid generation system also produced 81% of renewable energy source of all the total energy generated.**

Keywords—renewable energy; hybrid system generation; hotels; net present costs; cost of energy; wind turbine

I. INTRODUCTION

The role of hotels in the economy is highly recognized today, as the existence of the hotel can provide an important effect for regional development. Aside from being a local tourism accommodation service, the hotel has a significant role in the regional economy and contributes substantial foreign exchange to the state. The hotel operational itself mostly requires electrical energy and PLN becomes the main supplier of electrical energy in Indonesia. But PLN is a power supplier company that can not avoid the disturbance or damage that can disrupt the operation of the hotel. Hotel managers are trying to maintain the availability of electricity supply by adding a backup when the supply of electricity from the grid is off.

Diesel generator is a generator that is commonly used as a reserve power in the hotel, because it has some advantages that are: the starting time is fast and does not require a large area. However, the operational costs and fuel of diesel generators are the greatest among other conventional types of generators.

This study proposes the use of hybrid plant generator as a hotel reserve power generation. Hybrid plant generator is the right solution especially to solve the problem of energy production cost as hotel reserve. It could be said that hybrid system is more economical.

There are some software that can be used to facilitate the analysis process e.g. HOMER, Hybrid2, Matlab and so forth. And HOMER is used in this study [1].

Hybrid generatoration systems generally consist of two categories: pure hybrid power generation system and hybrid

plant with a combination of renewable and non-renewable energy such as diesel generators. The variation of renewable and non-renewable energy applied in this research is a combination of wind turbine and diesel energy. Where renewable energy is suitable in coastal areas, considering the location where the research conducted was in coastal area.

II. BASIC THEORY

A. HOMER PRO

HOMER is a software model developed by the US National Renewable Energy Laboratory (NREL) for the purpose of power plant optimization system. HOMER is equipped with an estimated output of system specification/capacity, lifecycle cost, and greenhouse gas emissions. The microgrid HOMER software provides the detailed chronological simulations and optimizations in a relatively simple and easy-to-use model. It is tailored to a wide range of projects. For the village power system or power system scale, HOMER can be used for two factors, namely the technical and economic part of the project being worked on. For larger systems, HOMER can provide an important view that compares the costs and feasibility of different configurations, so the designers can use software that is more specialized for technical performance models. HOMER sensitivity analysis helps determine the potential impact of uncertain factors such as fuel prices or wind speeds on a particular system.

B. Hybrid System Generation

Hybrid plant generation is a combination of two or more plants that generally consist of more than one generation of renewable energy and a conventional plant. The purpose of hybrid plant design is to obtain fuel efficiency (saving fuel) by the conventional plant (generally diesel generators) or increase the reliability of the plant with renewable energy sources. The current use of renewable energy is very high, now it is known that the PV and wind turbine (WT) has been widely applied as a plant. The natural characteristic of renewable energy such as PV and WT that can only produce energy at a certain time, make the basic application of hybrid generating systems.

C. Diesel Generator

The diesel generator is a conventional type of power plant powered by piston engine technology fueled by the High Speed Diesel (HSD). Diesel generators convert the motion energy into electrical energy. Low cost capital and high fuel efficiency can be obtained when it operate at maximum capacity. The disavantage of diesel generators is the expensive fuel cost as well as generating emissions gases that can damage the environment [3].

D. Wind Turbine

Wind turbines are windmills that use the principle of energy conversion, that is by converting the kinetic energy generated by wind into electrical energy [6]. The working principle of wind turbine, namely:

- The wind will blow and pound the fan blade.
- The rotation center raises torque on the fan blade.
- The other blades experience the same thing and there will be a rotation.
- The shaft is connected to the gearbox, the shaft rotational speed in the gearbox is increased by adjusting the gears ratio in the gearbox.
- The gearbox is connected to the generator. The generator converts mechanical energy into electrical energy, and the electricity flows from generator to transformer to raise the voltage and then distributes it to consumers.

In determining the type of wind turbine to be used in the research, it is important to know the power generated by the wind turbine. The resulting power is calculated using (1) [7].

$$Power\ (P) = {}^{1}/_{2} * \rho * V^{3} * A \quad (1)$$

Where :

P = Power (Watt)

ρ = Wind density (1.225 kg / m^3)

V = Wind speed (m / s)

A = Area of cross section (m^2) = 1/4 * π * D

D = cross-sectional diameter / wind turbine (m)

E. Net Present Cost

Net Present Cost (NPC) is the total cost of all installation and operational costs during the duration of the project. Net Present Cost (NPC) can be calculated using (2)

$$NPC = Capital\ Costs + Replacement\ costs + O\&M\ costs + Fuel\ Costs - Salvage \quad (2)$$

Where;

Capital Costs	= component capital costs
Replacement Costs	= component replacement cost.
O & M Costs	= operational and maintenance costs.
Fuel Costs	= fuel costs.
Salvage	= the remaining cost of the component

F. Cost of Energy

Cost of Energy (COE) is the cost required to generate every 1 kWh of electrical energy i.e. the result of annual costs divided by annual energy production on the hybrid plants. The COE value of each scenario uses (3).

$$COE = \frac{TAC}{E_{retserved}} \quad (3)$$

Where :

$E_{(tot.served)}$ = total annual energy used to serve the load (kWh)

TAC = total annualized cost incurred for hybrid generation.

G. Renewable Fraction

Renewable Fraction is a percentage of electrical energy derived from renewable energy sources from the total energy generated by the hybrid generation. The electrical energy from renewable energy sources is generated by Renewable Energy System (RES). So it can be calculated using (4).

$$RP = \frac{E_{tot.RES}}{E_{tot.sys}} * 100\% \quad (4)$$

Where;

$E_{tot.RES}$ = total electrical energy generated by renewable energy system (kWh)

$E_{tot.sys}$ = the total electrical energy generated by the generating system (kWh).

H. Total Energy Production

Total energy production is the total amount of electrical energy produced by all components that work during the operational period. It can be calculated using (5).

$$E_{tot.prod} = E_{gen.diesel} + E_{wind\ turbine} + E_{grid} \quad (5)$$

Where :

$E_{(tot.prod)}$ = total energy production (kWh)

$E_{(gen.diesel)}$ = total energy production of diesel generator (kWh).

$E_{(wind\ turbine)}$ = total energy production of wind turbine (kWh)

E_{grid} = total energy production from the PLN grid (kWh)

III. Research Methodology

A. Tools and materials

The research tools used in this study consist of hardware and software. The hardware used is a laptop with 2.5 GHz i3 core processor, 2GB of RAM, 1 GB of VGA and A4 paper. While the software used is HOMER PRO 3.8.6.

The research materials used consist of:

- the survey data conducted on the hotel
- blackouts data from PT PLN (Persero)
- geographical location through Google Maps website,
- Potential data of natural resources (wind) from the Meteorology Climatology and Geophysics Agency of Yogyakarta
- Some journals and articles on hybrid generation systems, and
- Online sales website of generating components.

B. Research Methodology

The first thing to do is to start the research, The next step is to collect any data related to peneletian. After getting the desired data, the next step is to do the simulation or modeling using HOMER software. After modeling, the modeled data is analyzed. Furthermore, the final step is to write the results of research reports and provide conclusions from the research conducted. This research can be described in Fig. 1.

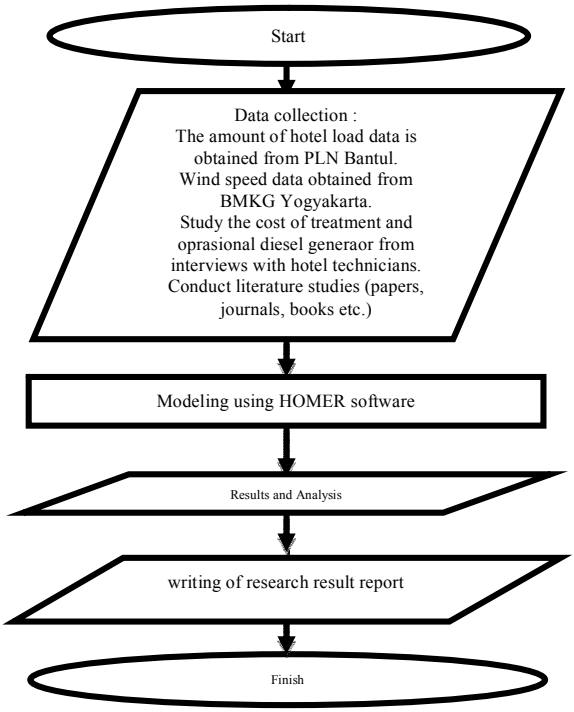


Fig.1. Research Methodology.

C. The main components of the system

The main components in doing this research consist of diesel generators, wind turbines, batteries and converters. For the main component of its own system is divided into 2 scenarios. Scenario 1 uses grid components and diesel generators, shown in Fig. 2. and scenario 2 using grid components, diesel generators, wind turbines, inverters and batteries can be seen in Fig. 3.

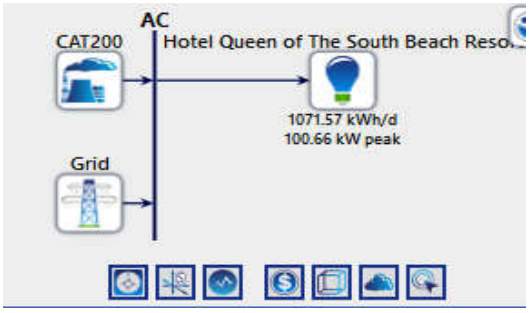


Fig. 2. The Schematic of 1st Scenario

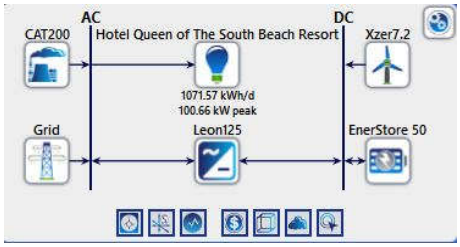


Fig. 3. The Schematic of 2st Scenario

IV. Results and Discussion

A. *Results of HOMER Software System Optimization*

The optimization process is conducted after the simulation process by the HOMER software is done. This simulation process aims to determine the optimum size of the variables on each of the main components installed during operation. The result of component optimization in scenario 1 consists of PLN Grid and 1 set of Caterpillar C9 200 kW diesel generator, can be seen in Fig. 4.

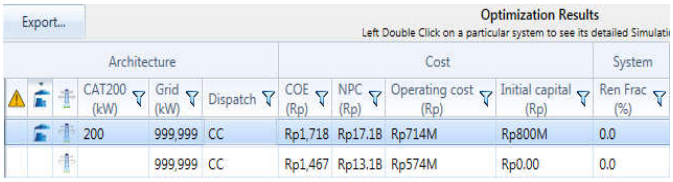


Fig. 4. Optimum System Configuration in Scenario 1, Results of HOMER Optimization

Scenario 2 consists of Grid PLN, 1 set Caterpillar C9 200 kW diesel generator, 7 units of Xzeres 7.2 10 kW wind turbine, 1 unit of EnerStore 50 Agile Flow battery and 1 unit Leonics GTP-507 125kW Converter, can be seen in Fig. 5.

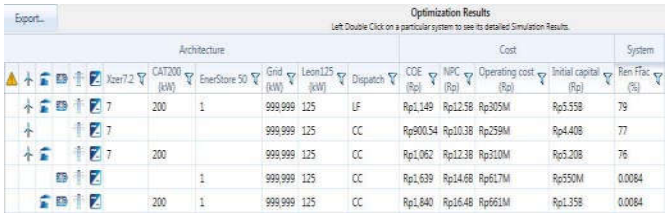


Fig. 5. Optimum System Configuration in Scenario 2, Results of HOMER Optimization

B. *Economic Value of Generation*

This study is divided into 2 scenario :

- Scenario 1

The economic value of a conventional generating system using a diesel generator is shown in Table I.

TABLE I. The Economil Value of the Reserve Generation Uses Scenario 1

Criteria	Value
Total Energy Production (kWh) / year	372.472
NPC (Rupiah)	Rp 15.952.360.144
Cost of Energy (Rupiah)	1.718
Renewable Fraction (%)	0

The total energy production using a diesel generator system is 372,472 kWh/year. The results of total energy production can be seen in Table II and calculated using (2)

TABLE II. The Total Annual Energy Production

Component	Production (kWh)
Generator	63.750
Grid PLN	308.722

$$E_{totprod} = 63.750 + 308.722\ kWh$$
$$E_{totprod} = 372.472\ kWh$$

The Net Present Cost is Rp 15,952,360,144. Net Present Cost results can be seen in Table III and calculated using (3).

TABLE III. Net Present Cost Result by HOMER Calculation

Component	*Capital* (Rp)	*Replacement* (Rp)	O&M (Rp)	*Fuel* (Rp)	*Salvage* (Rp)	Total (Rp)
CAT C9 ATAAC 200 kW	Rp 800.000.000	Rp 0	Rp 96.904.247	Rp 3.925.881.760	Rp 194.966.267	Rp 4.627.819.740
PLN	Rp 0	Rp 0	Rp 11.324.540.404	Rp 0	Rp 0	Rp 11.324.540.404
System	Rp 800.000.000	Rp 0	Rp 11.421.444.651	Rp 3.925.881.760	Rp 194.966.267	Rp 15.952.360.144

$$NPC = Rp\ 800.000.000 + Rp\ 0 + Rp\ 11.421.444.651 + Rp\ 3.925.881.760 - Rp\ 194$$
$$NPC = Rp\ 15.952.360.144$$

The Cost of Energy system is Rp 1.718/kWh. Total energy data and cost per year used to calculate Cost of Energy can be seen in Table IV and calculated using (4)

TABLE IV. TOTAL ENERGY DATA AND EXPENSES PER YEAR

Criteria	Value
Total Energy (kWh/year)	436.222
Annual Cost (Rp)	749.487.051

$$COE = \frac{\text{Rp } 749.487.051}{436.222\text{kWh}}$$
$$COE = \text{Rp } 1.718/\text{kWh}$$

with Renewable Fraction of 0%. The result of Renewable Fraction can be calculated by using (5).

$$RF = \frac{0}{436.222\ kWh} * 100\%$$
$$RF = 0\%$$

- Scenario II

The economical value of a hybrid wind-diesel reserve generation is shown in Table V.

TABLE V. THE ECONOMIC VALUE OF THE RESERVE GENERATION USING SCENARIO II

Criteria	Value
Total Energy Production (kWh)/year	491.562
NPC (Rupiah)	12.481.507.750
Cost of Energy (Rupiah)	1.148
Renewable Fraction (%)	81

Total energy production using hybrid wind-diesel system is 491,562 kWh/year. The results of total energy production can be seen in Table VI and calculated using (2).

TABLE VI. TOTAL ENERGY PRODUCTION PER YEAR

Component	Energy Production (kWh)
Generator	8.500
Wind Turbine	398.685
Grid PLN	84.377

$$E_{totpred} = 8.500\ kWh + 398.685\ kWh + 84.377\ kWh$$
$$E_{totpred} = 491.562\ kWh$$

The Net Present Cost is Rp 12.481.507.750. Net Present Cost results can be seen in Table VII and calculated using (2)

TABLE VII. THE NET PRESENT COST-HOMER CALCULATION RESULT

Component	*Capital* (Rp)	*Replacement* (Rp)	O&M (Rp)	*Fuel* (Rp)	*Salvage* (Rp)	Total (Rp)
CAT C9 ATAAC 200 kW	Rp 800.000.000	RRp 0	Rp 96.904.247	Rp 1.372.231.971	Rp 194.966.267	Rp 2.074.169.951
PLN	Rp 0	RRp 0	Rp 3.095.117.114	Rp 0	Rp 0	Rp 3.095.117.114
EEnerStore 50 Agile Flow Battery	Rp 350.000.000	Rp 628.880.165	Rp 114.004.996	Rp 0	Rp 146.224.700	Rp 946.660.461
Lleonics GTP-507 125 kW	Rp 200.000.000	Rp 359.360.094	Rp 114.004.996	Rp 0	Rp 83.566.971	Rp 589.808.119
Xxzeres 7.2 10 kW	Rp 4.200.000.000	Rp 3.637.771.729	Rp 570.024.982	Rp 0	Rp 2.632.044.607	Rp 5.775.752.105

$$NPC = Rp\ 5.550.000.000 + Rp\ 4.626.011.989 + Rp\ 3.990.056.335 + Rp\ 1.372.231.9$$
$$NPC = \text{Rp } 12.481.507.750$$

The Cost of Energy (CoE) system is Rp 1.148/kWh. Data of total energy and cost per year used to calculate Cost of Energy can be seen in Table VIII and calculated using (4)

TABLE VIII. TOTAL ENERGY SERVING THE LOAD AND TOTAL EXPENSES PER YEAR

Critria	Value
Total Energy Serving the Load (kWh/year)	476.920
Total expenses per year (Rp)	547.942.918

$$COE = \frac{\text{Rp } 547.942.918}{476.920\text{kWh}}$$
$$COE = \text{Rp } 1.148/\text{kWh}$$

with Renewable Fraction of 81%. The result of Renewable Fraction can be calculated by using (5)

$$RF = \frac{398.685}{491.562\ kWh} * 100\%$$
$$RF = 81\ \%$$

- Comparison of Economic Value of Hyu Wind-Diesel System Generator with Diesel Generator

The economical value of a wind-diesel hybrid system is compared to the economic value of a diesel generator as a hotel reserve generation. The comparison results are then used to determine the more economical backup system generators. Table IX shows the comparison of the economic value of a hybrid generator plant with a diesel generator.

TABLE IX. COMPARISON OF THE ECONOMIC VALUE OF HYBRID GENERATING SYSTEMS WITH DIESEL GENERATORS.

Criteria	Reserve Generation System	
	Scenario 1	Scenario 2
Total Energy Production (Kwh)	372.472	491.562
NPC (Rupiah)	Rp 15.952.360.1	12.481.507.750
Cost of Energy (Rupiah)	1.718	1.148
Renewable Franction (%)	0	81

Table IX shows the total energy production in scenario 1 is 436,222 kWh and in scenario 2 is 422,778 kWh. The economic value of the hybrid wind-diesel system is more economical than using diesel generators, where NPC of the system is reduced by Rp 2,577,352,064 or 15.08% and COE value which also decreased by Rp 150 or 8.73%. The wind-diesel hybrid generation system generator is also known to produce 67% of renewable energy source of the total energy generated

V. CONCLUSIONS

The economic study of the hybrid wind-diesel hotel reserve generation system has been done using the HOMER software. The Net Present Cost scenario 2 using hybrid wind-diesel system proved more economical by 21.75% compared to the reserve generation scenario 1 using diesel generator. Scenario 2 reserve generator that uses wind-diesel hybrid system has 33.17% lower Cost of Energy compared to the reserve generator scenario 1 that use diesel generator. The contribution of renewable energy sources to the use of hybrid wind-diesel system as a reserve generator of the hotel is able to generate renewable energy by 81% of the total energy produced. Through the economic study of reserve generation ystems at the Queen of The South Beach Resort Hotel using HOMER software shows the result that the wind-diesel reserve generation system proves to be more economical compared to using diesel generators

REFERENCES

[1] K. Daljeet and P. S. Cheema, "Software Tools for Analyzing the Hybrid Renewable Energy Sources:-A Review," *Int. Conf. Inven. Syst. Control*, pp. 1–4, 2016.

[2] L. Dl and L. Dengan, "SISTEM WIND-DIESEL UNTUK PEMBANGKIT ANGIN MENENGAH Dl INDONESIA." *LAPAN*, Vol.4,No.1,Juni, 2006.

[3] D. I. Pulau, and N. Karimunjawa, "Optimasi Sistem Pembangkit Listrik Hibrida Tenaga Surya , Jawa Tengah Dengan Menggunakan Perangkat Lunak," *Transient*, vol. 4, no. 4, pp. 10-29-1037, 2016.

[4] H. Mubarok, "Optimal specification analysis of hybrid-PV-battery-diesel power generation based on electrical outage cost as an industrial reserve power, " IEEE, vol. pp. 253 – 257, October 2017, International Seminar on Application for Technology of Information and Communication (isemantic) Semarang, 2017.

[5] M. H. Nehrir *et al.*, "A review of hybrid renewable/alternative energy systems for electric power generation: Configurations, control, and applications," *IEEE Trans. Sustain. Energy*, vol. 2, no. 4, pp. 392–403, 2011.

[6] R. N. Gunarto and E. Sarwono, "Study Eksperimental Berbagai Macam Jenis Sudu Turbin Angin Sumbu Horisontal Skala Laboratorium," *Tugas Akhir Universitas Sumatra Utara*, no. 111, pp. 1–11, 2016.

[7] T. Elektro, T. Elektro, and U. Brawijaya, "Studi analisis pembangkit listrik hybrid (diesel- angin) di pulau karimun jawa," *jurnal mahasiswa teub,* vol.2, no.5 pp. 1–8, 2014.

[8] A. K. Pradhan, S. K. Kar, and M. K. Mohanty, "Off-Grid Renewable Hybrid Power Generation System for a Public Health Centre in Rural Village," *Int. J. Renew. Energy Res.*, vol. 6, no. 1, pp. 282–288, 2016.

[9] M. Otong, A. Alimuddin, and I. Mas'ud, "Optimasi Kapasitas Pembangkit Listrik Tenaga Hibrida Menggunakan *HOMER* Di Pulau Tunda," *Setrum Sist. Kendali-Tenaga-Elektronika-Telekomunikasi-Komputer*, vol. 6, no. 1, pp. 1–13, 2017.

[10] KEMEN ESDM, "Peraturan Menteri Energi dan Sumber Daya Mineral Republik Indonesia No. 12 Tahun 2017 Tentang Pemnafaatan Sumber Energi Terbarukan Untuk Penyediaan Tenaga Listrik," pp. 2–22, 2017.

Classification of Cataract Slit-lamp Image Based on Machine Learning

Riyanto Sigit, S.T., M.Kom., Ph.D.[1)], Maulana Bayu Satmoko [2)], Dwi Kurnia Basuki, S.Si., M.Kom.[3)]
Informatic and Computer Departement
Electronic Engineering Polytechnic Institute of Surabaya
Surabaya-East Java, Indonesia
riyanto@pens.ac.id[1)], bsatmoko@gmail.com [2)], dwiki@pens.ac.id[3)]

***Abstract*— Cataracts are diseases caused by the presence of proteins in the lens that form abnormal and gradually enlarged clumps that will interfere with vision by blocking the light entering through the lens. Identification of cataracts is done by taking the image of the eye with a slit-lamp tool from the front of the eye. Slit-lamp images can provide information about the condition of the pupils that can only be analyzed by the doctor manually based on doctor's observation and doctor's experience that can cause different analysis in determining the actual eye condition. Things that are considered by the doctor in analyzing cataracts are the level of opacity in the eyes and the area covered by the turbid. Identification and classification with slit-lamp images can be performed better and more accurately using image processing techniques. Firstly, the grayscale method, median filter method and canny method is used to preprocess the slit-lamp images. Next, the hough circular method is used to automatically segment pupil from slit-lamp images. After the segmentation process, we use pixel scanning to extract mean intensity and uniformity from the pupil image. After the feature extraction process, classification is done by single perceptron based on the extracted feature. This research is expected to help the doctor to do cataracts classification so that the classification process will be easier and more accurate. Based on the test result show that the accuracy of the system is 96.6%.**

Keywords*— *cataract; segmentation; feature extraction; machine learning

I. INTRODUCTION

Cataracts are analyzed by the doctor through slit-lamp results. The analysis is done manually based on doctor's observation and experience. This makes it possible that doctor will have multiple analyzes about classification result, making it difficult to determine the actual state of the eye. Therefore we need a system that can analyze the pupil condition accurately to perform cataract classification. [1]

According to data from the Ministry of Health of the Republic of Indonesia, Indonesia is among the countries with the highest number of blind people due to cataracts in the world, especially Southeast Asia. The latest data, 1.5 percent per two million population are cataract patients and each year 240 thousand people are threatened with blindness. The number of people with cataract disease in Indonesia is as large as the number of people in Indonesia who are elderly in the year 2000 which is estimated about 7.4% of the total population in Indonesia or about 15.3 million people who experience cataracts. The number that has been dissected in 2015 is 4,425 eyes.

In the medical world, medical imaging technology has been widely used to obtain information about the human body. Medical imaging is expected to provide a better degree of accuracy for the identification and classification of cataract disease. Medical imaging generally includes preprocessing, image segmentation, feature extraction, identification and classification. The results of the classification will be used by the physician for both patient care and surgical considerations.

Examination Slit-lamp is one way to identify cataracts by knowing the condition of the pupils in the eye. Slit-lamp is a medical device that is able to see the state of the eye with lighting lamps. The lamps facilitate examination of the anterior segment or frontal structures and posterior segments of the human eye covering the eyelids, sclera, conjunctiva, iris, lens, and cornea. In general, cataract examination is done by Slit-lamp examination and continued with physician analysis [2] based on the image of Slit-lamp. As for the classification is done based on doctor's observation on the turbid area of the pupil.

This research will be done by identification and classification of cataracts with Slit-lamp image improvement step in order to provide optimal input on identification system and classification. Followed by Hough Circular method to find the location of pupils on the eyes and continued with feature extraction Mean Intensity and Uniformity on the image to do the classification by training on features that have been extracted using Neural Network Perceptron. It is expected that this system can be used to facilitate cataract classification system.

II. METHODOLOGY

To perform the classification of cataracts performed several stages including load image to insert the image slit-lamp on the system, preprocessing to improve image quality with binarization method, median filter and edge detection method, segmentation to get the location of the pupil and crop the pupil by using hough circular method, after the segmentation process, perform feature extraction to get some information from the pupil image such as intensity information and the opacity areas in the pupil area, doing machine learning using

the features available to get the weights that will be used to classify the cataracts into some classes.

Figure 1 is a procedure of the system used to classify cataracts with preprocessing stages, segmentation, feature-extraction and classification method.

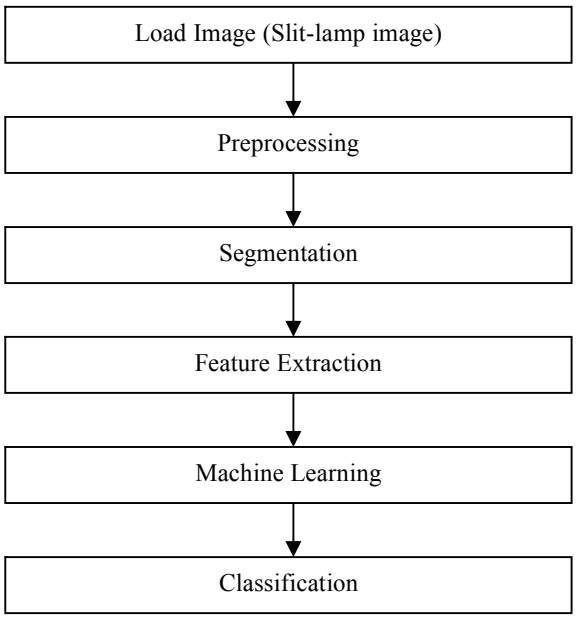


Fig. 1. System Diagram

A. Load Image

The input used in this system is a Slit-lamp image. Slit-lamp images will be obtained from Dr. Sjamsu Clinic in Surabaya. Image retrieval is done by observing light on the Slit-lamp device so that light does not affect the pupil area of the eye. Slit-Lamp light should be placed or positioned outside the pupil area by adjusting the image taking position. The image is taken with JPG extension and will be resized according to system requirement. Figure 2 is the result of a slit-lamp examination.

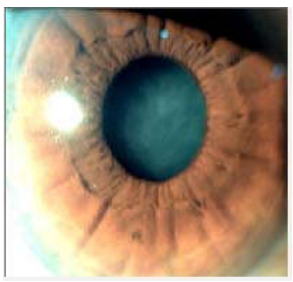


Fig. 2. Slit-lamp image

B. Preprocessing

Preprocessing is used to reduce noise, improve lighting effects and sharpen the edges of images. These are some preprocessing stages:

1) Grayscaling

The image input obtained from the Slit-lamp capture is an image with RGB specification. While to process the image in the segmentation and feature extraction required image with a specification of grayscale value 0-255 [3] so that in accordance with the needs of the next process. The grayscaling formula can be represented as :

$$Gray\ [i] = \frac{R[i] + G[i] + B[i]}{3}$$

Result of *grayscaling* [4] process can be seen in a Figure 3.

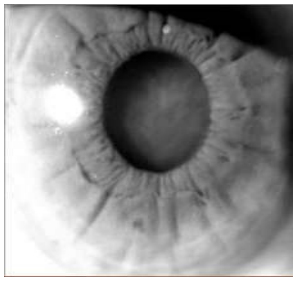


Fig. 3. Grayscale image

2) Median Filtering

Median Filtering is a non linier method that effective to reduce the noise. Median Filtering moving through pixel by pixel in an image, replacing each value with the median value of neighboring pixels. Median filtering is calculated by first sorting all pixel values from window to numerical sequence, and then replacing pixels with pixel middle values. The median filtering formula can be representated as:

$$f(x, y) = median_{(s,t) \in S_{xy}} \{g(s, t)\} \quad (1)$$

Result of *median filtering* process can be seen in a Figure 4.

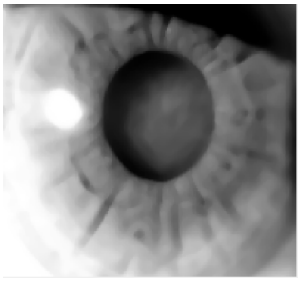


Fig. 4. Median filter image

C. Segmentation

Classification of cataracts can be done by considering the condition of pupils in the eye so that required segmentation process to separate the pupil with the whole image. This is preprocessing stage:

1) Hough Circular

Hough Transform is an image transformation technique that can be used to isolate or in other words acquire

features of an image. Since the purpose of a transformation is to acquire a more specific feature, Classical Hough Transform [5] is the most commonly used technique for detecting curved objects such as lines, circles, ellipses and parabolas. The main advantage of the Hough transform is that it can detect a edge with a gap at the feature boundary and quickly without being unaffected by noise. If the object is a circle, then the Hough circle is used. The procedure used in detecting a circle is the same as the Hough transformation of the line object [6], but it is done on a more complex dimensional space that is in the 3D space parameter (Xo, Yo, R). Where Xo and Yo are the central coordinates of the circle and r is the radius of the circle. Here is the equation of the hough circular method:

$$(x - x_0)^2 + (y - y_0)^2 = r^2 \quad (2)$$

Result of segmentation [7] process can be seen in a Figure 5.

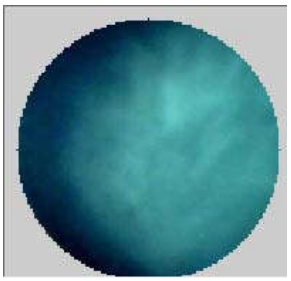


Fig. 5. Pupil image

D. Feature Extraction

Feature extraction is performed to find the mean intensity and uniformity values in the pupil image. These features are obtained by looking at the histogram value of the pupil area. The histogram used is a gray histogram where the gray histogram contains the gray pixel value in an image that has a range of 0-255 on each pixel. Here are the features that will be extracted on feature extraction [8]:

1) Mean Intensity

Mean intensity is the average intensity of gray values present in the pupil area. Mean intensity can be calculated by summing all the gray values in the pupil area then divided by the number of pixels in the pupil area like the following mathematical equations:

$$m = \sum_{i=0}^{L-1} i \left(\frac{I(i)}{N} \right) \quad (3)$$

Here 'm' is the average intensity, 'I' is the possible intensity, 'N' is the number of pixels in the image, and 'L' is the value of the possible intensity level. Mean Intensity indicates a turbidity area in the pupil or indicates cataract.

2) Uniformity

Uniformity is the value of the similarity of pixels present in the pupil. Uniformity can be calculated using the same gray value in one area like the following mathematical equations:

$$U = \sum_{i=0}^{L-1} \left(\frac{H(i)}{N} \right)^2 \quad (4)$$

Here 'U' is a measure of uniformity, 'H' is the probability histogram of the intensity level in the drawing region, and 'N' is the number of pixels in the drawing area, the index used is the index 0 to L - 1 into the corresponding histogram, and 'L 'is the value of the possible intensity.

Result of feature extraction process can be seen in a Figure 6.

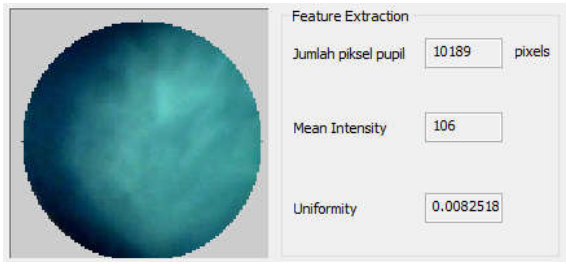


Fig. 5. Feature Extraction

E. Machine Learning

Machine learning is used to classify cataracts into several classes. Machine learning is done by using training data that has been classified by the doctor as a reference in doing cataract classification. Here are the method used to do system training:

1) Single Perceptron

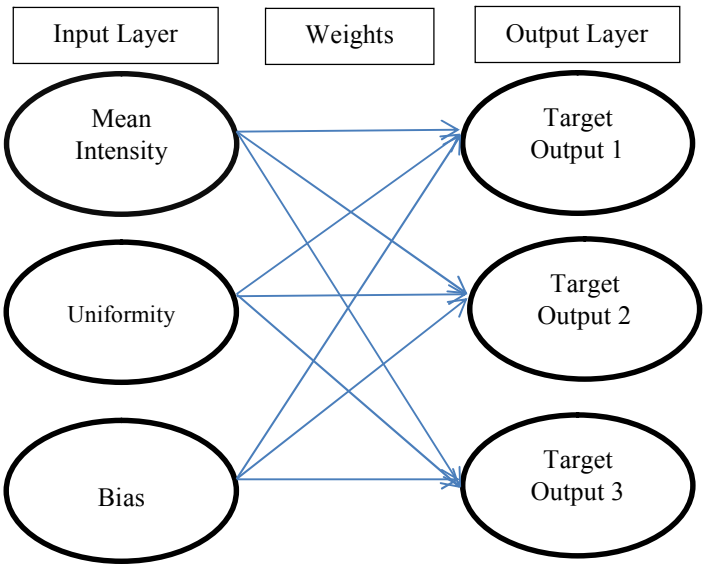


Fig. 6. Single Perceptron Training Model

In Figure 6, it can be seen that there are 9 weights that will be trained based on mean intensity, uniformity and bias. This training process will produce 3 outputs that match the targets that have been determined in the training process. The training process will succeed if the error value in the training process reaches 0. These trained weights will be used to classify independently in the classification process. In perceptron training there is normalization done so that the mean intensity and uniformity feature is worth between 0-1. The target of the training process is normal (1.0.0), mild-cataract (0,1,0), severe-cataract (0,0,1).

F. Classification

Classification is done by entering data testing on the classification system that has been established in the previous training system. The results of the classification include normal, mild, severe and unidentified / unknown. Based on the features of training data drawn from Dr. Sjamsu Clinic in September-October 2017 and based on figure 6 can be seen that the difference between normal eye, mild cataract, severe cataract is in pupil color. In image processing, the average value of intensity will be different for each class. In normal eyes the intensity will tend to be small due to the color on the pupils which tend to be dark and the uniformity value of pixels / uniformity will tend to close to 1 because the pixels in the eye are almost uniform at the same intensity value. In the eyes of mild cataract and severe cataract the intensity value will tend to be larger because there is a lighter intensity on the pupil area by the turbidity area of the pupil and the pixel uniformity value will be smaller because the intensity value contained in the pupil tends to have different values. The mean value of intensity and uniformity of each class can be separated so well that no multi-layer is required in the classification model. So mean intensity and uniformity can be written as follows [9]:

Mean Intensity: Normal <Mild <Severe
Uniformity: Normal> Mild> Severe

Neural network single layer perceptron can provide fast learning time but requires separate feature distribution between one class and another or there are no overlapping features.

G. Graphical User Interface

Graphical User Interface is used to facilitate medical personnel in using cataract classification system that will be done at Dr. Sjamsu Clinic. Here is a graphical user interface used for cataract classification in this research:

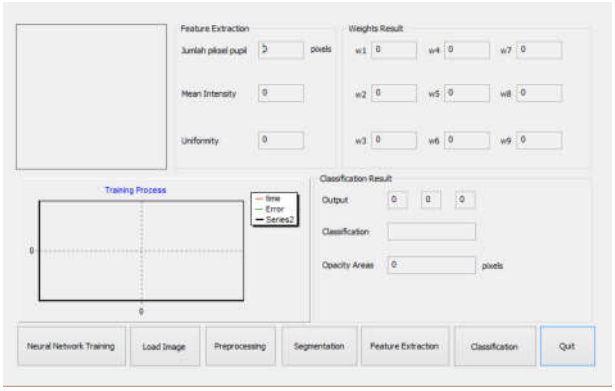
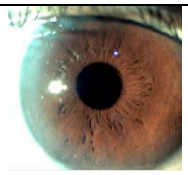
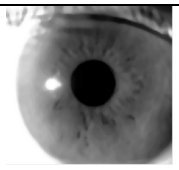
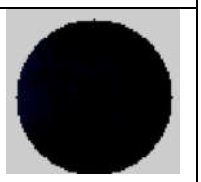
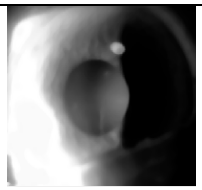
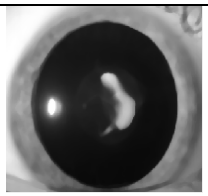
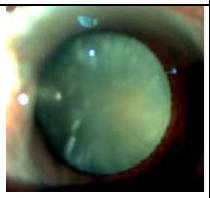
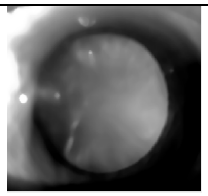
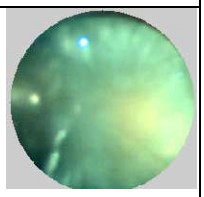
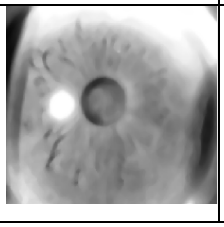
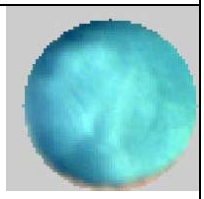


Fig. 7. Graphical User Interface

III. EXPERIMENTAL RESULT

In this section, will be explained about the final result of the system. Input data gathered from Dr. Sjamsu Clinic Surabaya. Table I show several data training preprocessing and segmentation result.

TABLE I. Preprocessing and Segmentation Result

No	Input Image	Preprocessing	Segmentation
1			
2			
3			
4			
5			

From table 1 it can be seen that preprocessing has been able to improve the image quality and segmentation process has been able to take the pupil part of the whole image.

Table II show data training feature extraction result. Here is a weight training on Neural Network Perceptron with 11 data (5 Normal, 3 Mild, 3 Severe) with the following features:

TABLE II. Feature Extraction Result

No.	Mean Intensity	Uniformity
1.	5	0.52
2.	6	0.42
3.	0	0.78
4.	4	0.87
5	0	0.98
6	126	0.008
7	55	0.016
8	125	0.0103
9	161	0.0065
10	183	0.00105
11	203	0.00136

Based on the features contained in table 2, the training process is done by Single Perceptron method to produce trained weights that are used for classification. After the training is done, the error graph appears as in Figure 8 below:

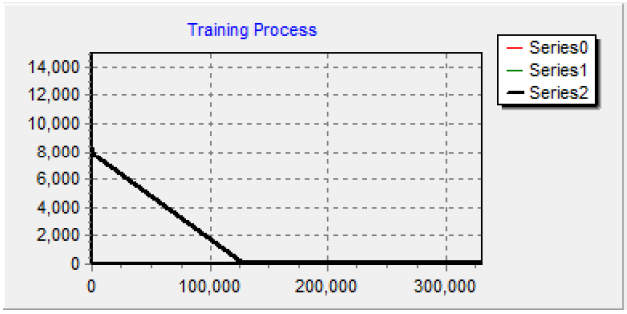


Fig. 8. Training Error Graph

Based on the graph in picture 8 the training process has been successfully done because the error in the training process has reached 0.

Table III shows the results of classification on 30 data. Classification is done by referring to the previous training process.

TABLE III. Feature Extraction Result

No	Medical Result	System's Output
1	Normal	Normal
2	Normal	Normal
3	Normal	Normal
4	Normal	Normal
5	Normal	Normal
6	Normal	Normal
7	Normal	Normal
8	Normal	Normal
9	Normal	Normal
10	Normal	Normal
11	Mild	Mild
12	Mild	Mild
13	Mild	Mild
14	Mild	Mild
15	Mild	Mild
16	Mild	Mild
17	Mild	Mild
18	Mild	Mild
19	Mild	Unidentified
20	Mild	Mild
21	Severe	Severe
22	Severe	Severe
23	Severe	Severe
24	Severe	Severe
25	Severe	Severe
26	Severe	Severe
27	Severe	Severe
28	Severe	Severe
29	Severe	Severe
30	Severe	Severe

Based on the 30 data testing, this system can classify 29 of 30 data correctly. It means that this system has 96.6 % of accuracy on 30 data test.

IV. CONCLUSION

The purposed of making this system is to do a cataract classification. Preprocessing has been successful in improving the image and segmentation has been successful in taking pupils from the entire image. The features of the pupils have been extracted and the classification training has been done successfully. The classification has been done on 30 data and has 3.33% of errors.

V. FUTURE WORK

This research can be improved by adding features that affect the classification of cataracts but the addition of features may cause overlapping features so that it will require additional layers on the neural network perceptron.

ACKNOWLEDGMENT

The authors wish to thank Politeknik Elektronika Negeri Surabaya that has facilitate this research and Dr.Sjamsu Clinic that has given medical information for many useful discussion and also Mr. Riyanto Sigit and Mr. Dwi Kurnia Basuki.

REFERENCES

[1] Amol B. Jagadale, D. V. Jadhav, 2014, Early Detection and Categorization of Cataract using Slit-Lamp Images by Hough Circular Transform, International Conference on Communication and Signal Processing, April 6-8, 2016, India.

[2] Optometric Clinical Practice Guideline, *Optometric Clinical Practice Guideline Care of the Adult Patient with Cataract*, 2010.

[3] Dmitry Nikolaev, COLOR-TO-GRAYSCALE IMAGE TRANSFORMATION PRESERVING THE GRADIENT STRUCTURE, Institute for Information Transmission Problems, Russia.

[4] Mohamed Kamel, Aiguo Zhao, 1992, Extraction of Binary Character/Graphics Images from Grayscale Document Images, CVGIP Vol. 55, No. 3, 1993, Canada.

[5] Alzbeta Vlachynska1, Zuzana Kominkova Oplatkova1, Martin Sramka, 2015 , The Coordinate System of the Eye in Cataract Surgery: Performance Comparison of the Circle Hough Transform and Daugman's Algorithm, AIP Conference Proceedings, Czech Republic.

[6] Milad Soltany, Saeid Toosi Zadeh, Hamid-Reza Pourreza, 2011, Fast and Accurate Pupil Positioning Algorithm using Circular Hough Transform and Gray Projection, International Conference on Computer Communication and Management, 2011, Singapore.

[7] Ann A. Jarjes, 2011, Improved Greedy Snake Model for Detecting Accurate Pupil Contour, 2011 3rd International Conference on Advanced Computer Control, 2011, China.

[8] M. K. Mandal, and T. Aboulnasr, Fast Wavelet Histogram Techniques for Image Indexing, School of Information Technology and Engineering, Canada.

[9] Barraquer RI, Pinilla Cortés L, Allende MJ, Montenegro GA, Ivankovic B, D'Antin JC, Martínez Osorio H, Michael R, Validation of the Nuclear Cataract Grading System BCN 10, Ophthalmic Res. 2017.

Improving Classifier Performance Using Particle Swarm Optimization on Heart Disease Detection

[1]Sugeng Hendra Wijaya, [2]Gansar Timur Pamungkas, [3]M. Burhanis Sulthan, [4]Muljono
[1,2,3,4] Informatics Engineering Department
Dian Nuswantoro University
Semarang, Indonesia
[1]sugengagel@gmail.com, [2]gansar.pamungkas@gmail.com, [3]burhan.sulthan33@gmail.com, [4]muljono@dsn.dinus.ac.id

***Abstract*— Data mining is an evolving technology that aims to extract information from very large data to generate useful knowledge. Lots of datasets provided online can be beneficial in the context of research, such as medical datasets obtained from public and private repository developed to conduct a clinical diagnosis of diseases. Hence, we employ a public dataset from the UCI Machine Learning Repository on heart disease. in the current study, we proposed an improvement of the classifier, i.e. Naive Bayes, Decision tree, and k-Nearest Neighbor on detection of heart disease using the particle swarm optimization. additionally, the proposed approach runs to select some significant features as data entries for the classification phase in order to improve the performance of the classifiers. the Experimental results show that our built-model in selecting prime features increases the performance accuracy of heart disease detection. eventually, The main objective, as well as advantage of this study, are to provide a significant improvement of classifiers' performance, which then resulted in the correct diagnosis on time.**

Keywords— Data mining; Heart disease; Feature selection; Naive Bayes; Decision tree; k-NN; Particle Swarm Optimization.

I. INTRODUCTION

Nowadays, human health is a very important aspect of continuing their livings. Human lifestyles have a big impact on their health. Based on the statistical data from the WHO (World Health Organization) in 2012 stated that 31% of deaths in the world were caused by heart disease. This statistical data proves that an early education for people are required and prevention for typical patient's behavior leads to heart disease should be prioritized.

Based on the research of Babic F, et al [1] stated that HD (Heart Disease) is the name of conditions or disorders that occur in the heart or blood vessels. This disease can be caused due to various diseases including stroke, hypertensive, cardiomyopathy, coronary artery diseases, etc [2].

Early prevention of heart disease can reduce the number of death. Based on a medical dataset owned by a hospital or health clinic can be provided for clinical diagnosis to patients that can help doctors or general practitioners for identification. Expected result from the prediction of the heart disease can provide effective prevention and early detection of patients who get heart disease.

This paper consists of four main sections. At first, we introduce the dataset to support the effective diagnostics of heart disease. This paper uses public dataset retrieved from the UCI (University of California, Irvine) Machine Learning Repository. The second section describes the result of the classification phase in predicting the heart disease using 3 (three) algorithms are NB (Naive Bayes), k-NN (K-Nearest Neighbor), and decision tree. The third section discusses improving the accuracy of prediction using feature selection technique. Due to that feature selection, we have proposed a model of selecting features using PSO (Particle Swarm Optimization) technique. The last section concludes the whole discussion and states an evaluation and development for our future work.

A. Related Work

In this study, we investigated the use of data mining techniques purposed to detect heart disease. Generally, lots of factors affecting the heart disease such as blood, cholesterol, chest pain, sugar, etc. For that concern, these factors are used as the prime aspects in diagnosing the patients' heart disease. Chaitrali S. Dangare, et al [3] stated that some difficulties could happen during the analysis. On the one hand, Zeinab Arabasadi, et al [4] conducted an experiment to detect coronary artery disease (CAD) using the hybrid neural network-Genetic algorithm. On the other hand, Purushottama, et al [5] built a system for predicting heart disease using the rule-based method in data mining domain. This could easily be made if the data was in a small amount, yet so complicated in a huge amount of data, due to the time consuming and efficiency. Still, the heart disease kinds were coronary heart diseases (CHD), cardiovascular disease (CVD), and coronary arteries diseases (CAD), which were diagnosed in the research of [6] using several data mining techniques. Those resulted in the prediction dependent on the different classification methods.

Sarath Babu et al [7] surveyed a public dataset using data mining technique implemented in predicting the heart disease. The techniques and methods are a genetic algorithm, K-Mean, decision tree and MAFIA (MAximal Frequent Itemset Algorithm) algorithm, for the classification process. Based on the results of this paper, the decision tree achieved the best result after applying a genetic algorithm to choose the optimal features. [8] Performed an experiment of detecting heart disease

using PCA (Principle Component Analysis) feature selection technique and k-NN (k-nearest neighbor) classification algorithm. The experiment was also adapted the clustering technic, which covered the unseen data. The input (data entries) from the clustering technic was then processed to the built model. On the use of the technique, [9] implemented the binary version of PSO, i.e. BPSO (Binary Particle Swarm Optimization) incorporated with the C4.5 algorithm. While [10] built a model to diagnose dengue disease by implementing the incorporation of PSO and ANN (Artificial Neural Network).

Kanika Pahwa and Ravinder Kumar [11] in their study designed a hybrid model for diagnosing heart disease using Naive Bayes and Random Forest algorithm. The UCI Machine Learning dataset had been used to perform the experiment. For feature selection, this research used SVM-RFE and Gain Ratio algorithms that were applied to the dataset. This research result got an increased accuracy after using feature selection technique, the Random Forest from 76,8977% to 84,1604% and Naive Bayes from 78,2178% to 84,1584% respectively. Another research [12] identified the EHR (Electronic Health Record) heart disease dataset of a hospital in Boston using several classification algorithms. A research [13] used medical practitioners dataset in South Africa to predict heart disease and implemented some algorithms, i.e. J48, Naïve Bayes, REPTREE, Simple CART, and Bayes Net on that dataset. In these concern, those researches tried some methods and implemented those to either private and public dataset purposed for the effectiveness of heart disease prediction.

B. Dataset

For the attributes that exist, only a few portions are used for the classification process. Not all attributes are used, since attributes that do not have an important role in the classification will be discarded with feature selection techniques. 270 (two hundred seventy) records of existing data will be divided into two part phases namely training and testing. Training phase with 243 records of data is used to create a classification model, while testing phase with 27 records of data is used to evaluate the model built, purposed to know how accurate the model.

In this research, we use a dataset which was retrieved from the UCI (University of California, Irvine) Data Mining Repository that can be accessed in http://archive.ics.uci.edu/ml/machine-learning-databases/statlog/heart/. For the attributes used in the dataset are totally 13 attributes such as age, sex, chest pain type, etc wholly covered in 270 records of data.

II. RESEARCH METHOD

In this study, we used the method used for the classification process. The technique used in the feature selection process is PSO (Particle Swarm Optimization). The explanation of each method and technique used are listed below:

A. Particle Swarm Optimization

The PSO (Particle Swarm Optimization) is an effective optimization algorithm that solves the problems of the ANN (Artificial Neural Network) algorithm, which generally works through the Backpropagation algorithm. Furthermore, [14] stated that the PSO has more comparisons for feature selection and has superior performance for many optimization problems with faster and more stable convergence rates. Conceptually, Sweta Mishra and Yamini Chouhan [15] stated that the PSO performed the sharing of information between the particles that will assist in finding the optimal solution. The PSO search method runs through iteration, which lies on the foundation of learning the movement of fish and bird organisms [16].

The PSO is used to solve some optimization problems. Similar to the GA (Genetic Algorithm), Particle Swarm Optimization searches using populations (swarm) of the individuals or candidate solutions (particles) to be updated through iterations. Particle Swarm Optimization has several parameters such as position, acceleration constant, speed, maximum speed, and weight of inertia.

Assume that the PSO uses two processes to complete the mission. First, it stated the initial velocity value running in an iteration or the size of the movement of the particles to 0 (zero). On this initial particle movement, then determine the initial position through equation (1).

$$X = x_{min} + rdm\ [0,1]\ \text{x}\ (x_{max} - x_{min}) \tag{1}$$

then,

$$vel_{jur}^{(\text{ty}+1)} \leftarrow \alpha vel_{jur}^{(ty)} + Uni(0,\beta) \left(\text{par}_{jur} - \text{x}_{jur}^{ty}\right) + Uni(0,\beta)\left(\text{par}_{zur} - \text{x}_{jur}^{ty}\right) \tag{2}$$

In which the equation (2) can be clarified into j is the particle index of the target, ur is the data dimension, x_j is the position of the particle, vel_j is the particle's velocity, par_j is every position, which is found with j, z is the relation index of the particle on j, α and β are permanent on every iteration, and $Uni(0,\beta)$ is a random number randomly generated in iterations.

When the iteration of the PSO is running, the second process within the whole process is updating the velocity. Every particle cycle defined as the variable in the centroid are positioned-updated as well as updating the par_j and par_z. to update the position of best par_j and par_z in the iteration, the equation (3) is employed.

$$x_{jur}^{(ty+1)} \leftarrow x_{jur}^{ty} + vel_{jur}^{(\text{ty}+1)} \tag{3}$$

B. Naïve Bayes

Naïve Bayes is categorized as one of the classification algorithms that make the learning process easier by assuming the class features are provided independently [17], [18]. Therefore, each feature is explicitly not related. For example, a feature with a representation of the student's value is not related to the level of attendance in the class, even though these features are on the same domain. For instance, there exists a feature that does not have a class, then the feature will be ignored.

Naïve Bayes is a probabilistic approach that can be applied to the field of classification [19], [20], [21]. Naïve Bayes focuses on predicting the probability of words given to a particular class and looking for the greatest value [22]. The Naive Bayes algorithm records how often the value of the target field appears along with the input field value. It considers each symptom to contribute independently to the likelihood of a patient suffering from heart disease. This estimates the probability of observing a particular value in a particular class with its frequency ratio in the interest class over the previous frequency of the class. Bayes theory is written in the form of equations as follows [18]:

$$P(c)=\frac{Fc}{Fd} \quad (4)$$

In the equation (4), *Fc* is used to describe all training documents labeled with a *C* target variable. Then *Fd* is notated to describe all training documents.

$$P(c \mid d) = P(c)\prod_{i=1} nP(x_i \in d \mid c) \quad (5)$$

Equation (5) is used to calculate the probability of a document against each target variable.

C. k-Nearest Neighbor

Another method implemented in the classification phase is the k-NN (k-Nearest Neighbor) algorithm. It employs the label of the majority of its neighbors closest to *k* in the training venue to classify the unlabeled data object. Thus, its simplicity, often produce competitive results due to the incorporation of some other prior learning algorithms in some domains. Furthermore, It has an advancement in the current state [23] significantly. Other researches [24], [25], [2] implemented this method were also done recently.

The algorithm step in this study is to calculate the distance from the test point for all training points. Then, sort by a distance of the specified data points. After all, apply an intelligent calculation to the class label for each of the *K* nearest neighbors and choose the training point according to the label as a result of the prediction.

In this study, we employed three values of *k,* i.e. $k = 1$, $k = 3$, and $k = 5$ for the dimensional space as a prediction label from each new point that will be the label of the sample point in the middle of the cell itself. It's easy to verify that we can get rid of predictions with the majority of labels by the right choice. In general, in *M*-dimensional space, by taking the weighted average of the target according to the Euclidean distance measurement between the vector *q* and *r* [26] is given with the following equation:

$$m(q,r)\left[\sum_{m=1}^{M}(q_m - r_m)^2\right]^{\frac{1}{2}} \quad (6)$$

D. Decision Tree

It is said in the research of Gandhi M, et al [27] that a decision tree became one of the most popular algorithms on classification phase due to its easiness to be interpreted by a human. This lies beyond a concept that a prediction model, which uses a tree structure or a hierarchical structure could strongly be implemented to either small and huge data. It is to transform the data entries into decision rules and decision trees. It is said that one main benefit of using this algorithm is the ability to simplify some complex decision-making processes. Thus, the decision makers interpret the problem solving easily. As for the steps of the decision tree algorithm as follows:

Input:

- Data division, *S*, several training datasets, which relate to a class label.
- Feature list, candidate features in a set.
- Feature selection techniques, the process of determining the separation between the 'best' criteria from a training data related to an individual class and also the distribution of features from a criterion.

Output: Decision tree.

The measurement of choosing a feature in this algorithm is the gain information contained in a feature [28]. A node *T* is represented in a proportion of data *S*. The separator feature for node *T* is selected with the highest information gain from the existing features. Selected features minimize information on the number of tests to classify training data in a class and produce a simple decision tree. The information contained in a feature is needed to classify a data sharing in *S* calculated as follows:

$$Info(S) = -\sum_{i=1}^{m} bi \log_2(bi) \quad (7)$$

where *bi* is the probability that the proportion of random data in *S* includes the class *Ci*. The function used in the *log* is base 2, due to the coding of information is in the form of bits. The *info (S)* is the average information that is added up and will be needed for the process of identifying labels from the proportion of data in *S*. The *info (S)* is also called as *entropy S*.

III. Experimental Result

In this research, we compare the result of analysis of heart disease using Naive Bayes, Decision Tree, and k-NN algorithms along with the implementation of PSO feature selection technique. A design of the proposed model in our research is shown in Fig. 1 below.

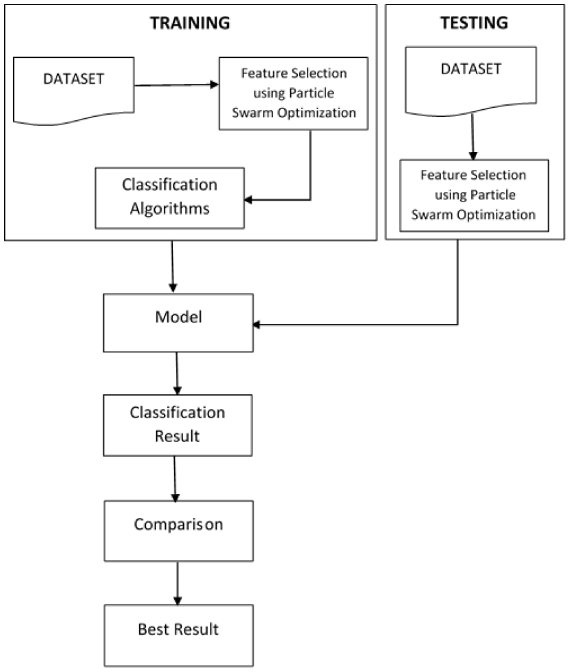


Fig 1. Design of the proposed model

Figure 1, shows that this research compares the results of classification with various algorithms tested before and after the feature selection using Particle Swarm Optimization. For the research evaluation, this paper uses the confusion matrix table, in which for instance the dataset consists of only two classes, one class is considered as positive and the other one is negative.

In this current research, we conducted the experiment using a PC (Personal Computer) with the following specifications: the Processor: Intel (R) Core (TM) i7-4700MQ, CPU 2.4 GHz. RAM (Random Access Memory) embedded: 4 (four) GB. The system type: 64-bit Operating System. For the VGA installed: NVidia GeForce GT745M, 2 (two) GB DDR3. The operating system running on this personal computer is Windows 7 (seven).

TABLE 1. CONFUSION MATRIX

	Predicted: Positive	**Predicted: Negative**
Actual : Positive	True positives	False positives
Actual: Negative	False positives	True positives

After the test data is entered into the table 1, the calculated values have been entered to obtain the amount of recall, precision, and accuracy. Here is the result of comparison result measurement on the dataset with unset feature selection technique as formed in table 2.

TABLE 2. RESULT COMPARISON

	Accuracy	**Precision**	**Recall**
Naive Bayes	84%	86%	80%
C.45	72,96%	76,40%	76%
k-NN (k=1)	57,78%	61%	64,67%
k-NN (k=3)	65,56%	68,74%	70%
k-NN (k=5)	65,19%	67,59%	72,67%

In addition to improving the results of the analysis, we perform feature selection using Particle Swarm Optimization (PSO) to get better features, Table 3 shows the result of the comparison:

TABLE 3. RESULT COMPARISON WITH PSO

	Accuracy	**Precision**	**Recall**
Naive Bayes	86,67%	87,77%	88,67%
C.45	80,74%	81,51%	85,33%
k-NN (k=1)	73,33%	73,36%	82%
k-NN (k=3)	82,96%	81,81%	90%
k-NN (k=5)	84,81%	83,07%	92%

Based on the results shown in Table 3, it is known that by using the Particle Swarm Optimization feature selection technique can particularly improve the accuracy of all algorithms.

The comparison of accuracy, precision, and recall results on the implementation of PSO feature selection technique can be seen below. The blue bar represents a performance before performing feature selection technique and the performance after performing feature selection technique is represented in the red bar.

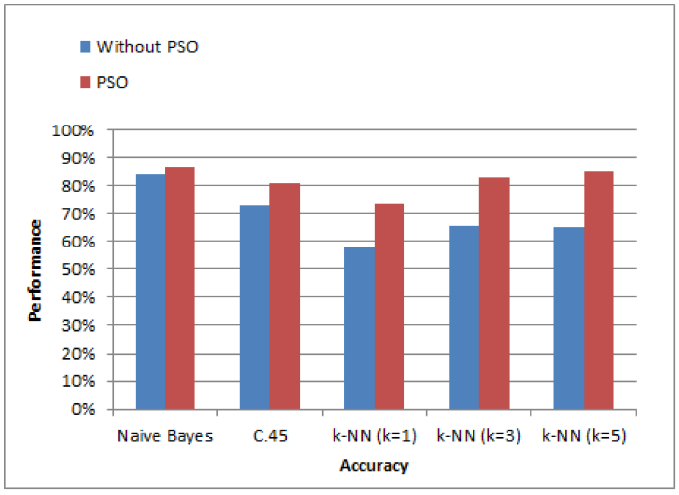


Fig 2. Graph of comparative accuracy rate before and after feature selection

Figure 2 shows that there is an increase in accuracy after implementing a feature selection technique. The Naive Bayes algorithm got the highest accuracy with 84% to 86.67% respectfully.

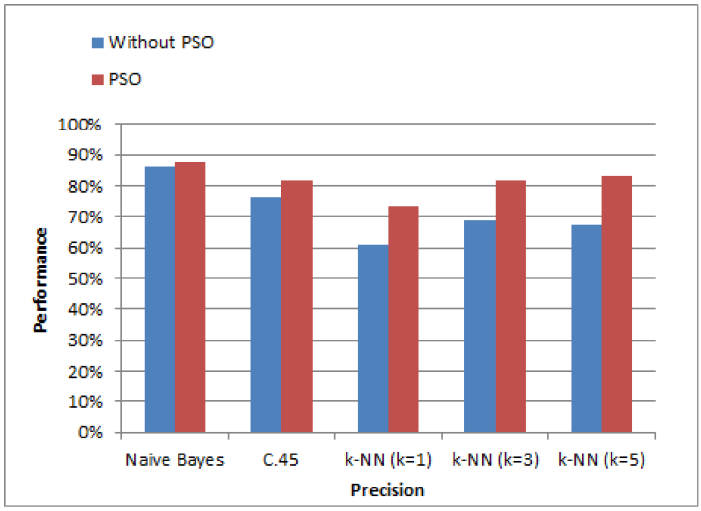


Fig 3. Graph of the comparative precision rate before and after feature selection

Figure 3 shows that there is an increase in precision after implementing a feature selection technique. The Naive Bayes algorithm got the highest precision with 86% to 87,77% respectfully.

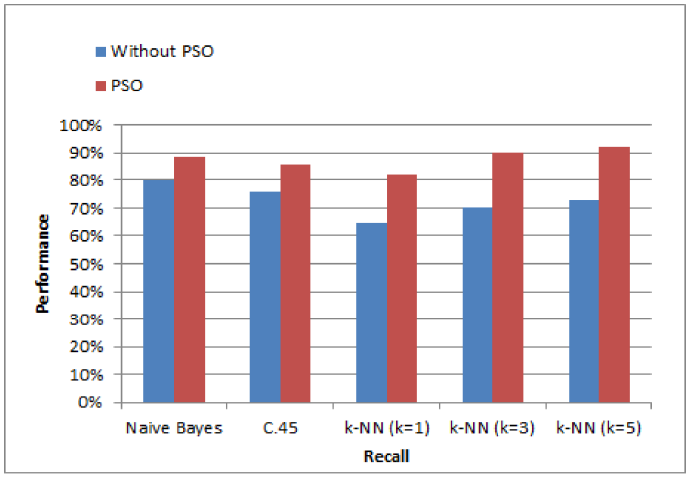


Fig 4. Graph of comparative recall rate before and after feature selection

As shown in figures (2) and (3), Naive Bayes has the best results of bathed with other algorithms with the highest accuracy of 86.67% and precision 87.77%. Naive Bayes algorithm achieved the best result after implementing the feature selection technique by eliminating some attributes. From the experimental results, it is known that the performance of classifiers is improved by implementing the Particle Swarm Optimization feature selection technique as obtained in table 4 below.

TABLE 4. RESULT OF FEATURE SELECTION USING PSO

Attribute	Weight
sex	1,000
restecg	1,000
thalach	1,000
oldpeak	0,875
ca	0,639
thal	0,530
cp	0,122
fbs	0,064
age	0,000
tresbps	0,000
chol	0,000
exang	0,000
slope	0,000

In Table 4 shows 13 (thirteen) features and after feature selection process resulting 8 (eight) attributes that are used such as sex, restecg, thalach, oldpeak, ca, thal, cp, and fbs. Other attributes are not used due to the low attribute weights are 0 (zero).

The calculation results are visualized with the ROC (Receiver Operating Characteristics) curve. The comparison of the two classes can be seen in Figure 5, which is the ROC curve for the Naïve Bayes and Particle Swarm Optimization algorithms.

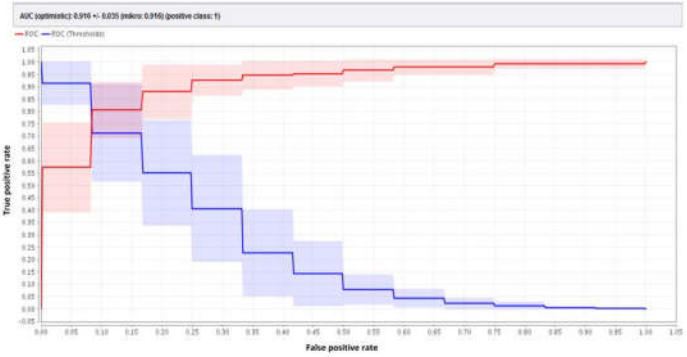


Fig 5. Graph of ROC Curve

The ROC (Receiver Operating Characteristics) curve is a technique used in research to visualize the performance results of a classifier. Generally, [29] stated that this graph is used in the medical field, but lately it is applied in the fields of

computing, data mining and machine learning.

Figure 5 shows the result of the ROC Curve with Naïve Bayes Algorithm and Particle Swarm Optimization. The accuracy results obtained by using the Naïve Bayes algorithm due to its high performance compared with other classification algorithms and Particle Swarm Optimization features selection is 86.67% and AUC (Area Under the Curve) value is 0.916.

IV. CONCLUSION

At the end of this research, stated the main purpose of this research was to classify patients with heart disease by using a public dataset obtained at UCI Machine Learning. In this paper, we compare the performance results using 3 (three) classification algorithms are Naive Bayes, Decision Tree, and k-NN. There is an evident from the results that the Naive Bayes algorithm performs well through achieving the highest accuracy value. To improve the accuracy, this experiment adds Particle Swarm Optimization as a feature selection technique. The result of the accuracy of Naive Bayes has increased from 84% (13 features) to 86.67% (8 features). Thus, the Particle Swarm Optimization feature selection technique can boost the classifiers' performance accuracy. From the classification results are expected to assist the Doctor or other health practitioners in conducting the correct heart disease detection in case of decision making so as to reduce the mortality rate caused by heart diseases.

REFERENCES

[1] F. Babič, J. Olejár, Z. Vantová, and J. Paralič, "Predictive and Descriptive Analysis for Heart Disease Diagnosis," *Proc. Fed. Conf. Comput. Sci. Inf. Syst.*, vol. 11, pp. 155–163, 2017.

[2] M. A. jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm," *Procedia Technol.*, vol. 10, pp. 85–94, 2013.

[3] C. S. Dangare and S. S. Apte, "a Data Mining Approach for Prediction of Heart Disease Using Neural Networks," *Int. J. Comput. Eng. Technol.*, vol. 3, no. 3, pp. 30–40, 2012.

[4] Z. Arabasadi, R. Alizadehsani, M. Roshanzamir, H. Moosaei, and A. A. Yarifard, "Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm," *Comput. Methods Programs Biomed.*, vol. 141, pp. 19–26, 2017.

[5] Purushottam, K. Saxena, and R. Sharma, "Efficient Heart Disease Prediction System," *Procedia Comput. Sci.*, vol. 85, pp. 962–969, 2016.

[6] S. M. Alzahani, A. Althopity, A. Alghamdi, B. Alshehri, and S. Aljuaid, "An Overview of Data Mining Techniques Applied for Heart Disease Diagnosis and Prediction," *Lect. Notes Inf. Theory*, vol. 2, no. 4, pp. 310–315, 2015.

[7] H. M. Sarath Babu, Vivek EM, Famina KP, Fida K, Aswathi P, Shanid M, "Heart Disease Diagnosis Using Data Mining Technique," in *International Conference on Electronics, Communication and Aerospace Technology*, 2017, pp. 750–753.

[8] S. Gupta, P. A. Nema, and P. K. Agrawal, "Heart Disease Prediction using PCA-KNN in Data Mining."

[9] L. Brezocnik, "Feature selection for classification using particle swarm optimization," no. July, pp. 6–8, 2017.

[10] S. Gambhir, S. K. Malik, and Y. Kumar, "PSO-ANN based diagnostic model for the early detection of dengue disease," *New Horizons Transl. Med.*, vol. 4, no. 1–4, pp. 1–8, 2017.

[11] K. Pahwa and R. Kumar, "Prediction of heart disease using hybrid technique for selecting features," *2017 4th IEEE Uttar Pradesh Sect. Int. Conf. Electr. Comput. Electron.*, pp. 500–504, 2017.

[12] W. Dai, T. S. Brisimi, W. G. Adams, T. Mela, V. Saligrama, and I. C. Paschalidis, "Prediction of hospitalization due to heart diseases by supervised learning methods," *Int. J. Med. Inform.*, vol. 84, no. 3, pp. 189–197, 2015.

[13] H. D. Masethe and M. A. Masethe, "Prediction of Heart Disease using Classification Algorithms," *Proc. World Congr. Eng. Comput. Sci.*, vol. II, pp. 22–24, 2014.

[14] M. Kaur, K. Singh, and N. Sharma, "Data Mining as a tool to Predict the Churn Behaviour among Indian bank customers," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 2, no. May, pp. 720 – 725, 2013.

[15] Y. C. Sweta Mishra, "A Review on PSO and Association Rule Mining for Item Set Generation," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 5, no. 2, pp. 1302–1309, 2017.

[16] G. I. Webb, "Posterior Probability," *Encycl. Mach. Learn.*, p. 780, 2010.

[17] E. D. Michie, D. J. Spiegelhalter, and C. C. Taylor, *Machine Learning , Neural and Statistical Classification*, vol. 37, no. 4. 1994.

[18] R. Wongso, F. A. Luwinda, B. C. Trisnajaya, O. Rusli, and Rudy, "News Article Text Classification in Indonesian Language," *Procedia Comput. Sci.*, vol. 116, pp. 137–143, 2017.

[19] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," *AAAI/ICML-98 Work. Learn. Text Categ.*, pp. 41–48, 1998.

[20] C. Catal and M. Nangir, "A sentiment classification model based on multiple classifiers," *Appl. Soft Comput. J.*, vol. 50, pp. 135–141, 2017.

[21] J. D. M. Rennie, L. Shih, J. Teevan, and D. R. Karger, "Tackling the Poor Assumptions of Naive Bayes Text Classifiers," *Proc. Twent. Int. Conf. Mach. Learn.*, vol. 20, no. 1973, pp. 616–623, 2003.

[22] M. Vadivukarassi, N. Puviarasan, and P. Aruna, "Sentimental Analysis of Tweets Using Naive Bayes Algorithm," vol. 35, no. 1, pp. 54–59, 2017.

[23] Y. Udovychenko, A. Popov, and I. Chaikovsky, "K-NN binary classification of heart failures using myocardial current density distribution maps," *2015 Signal Process. Symp. SPSympo 2015*, no. Cddm, 2015.

[24] M. Muljono, M. Nurul, S. Winarsih, and C. Supriyanto, "Evaluation of Classifier for Indonesian Text Emotion Detection," *Int. Semin. Appl. Technol. Inf. Commun.*, pp. 130–133, 2016.

[25] M. Shouman, T. Turner, and R. Stocker, "Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients," *Int. J. Inf. Educ. Technol.*, vol. 2, no. 3, pp. 220–223, 2012.

[26] H. Daume, *A course in machine learning*, Second edi. Self-published, 2012.

[27] M. Gandhi, "Predictions in Heart Disease Using Techniques of Data Mining," *Int. Conf. Futur. trend Comput. Anal. Knowl. Manag.*, 2015.

[28] H. J. and K. Micheline, *Data Mining: Concept and Techniques*, Second edi. Morgan Kaufmann, 2006.

[29] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

Improved Edge Detection Based on Adaptive Gaussian Smoothing in X-Ray Image

Derry Setiawan[1], Ade Pratama[2], Sudibyo[3], MA Soeleman[4]

[1,2,3,4] Department of Informatics Dian Nuswantoro University

Imam Bonjol Street 207, Pindrikan Lor

Semarang Tengah Indonesia

[1]1111206944@mhs.dinus.ac.id

[4] Arief22208@gmail.com

Abstract—*Currently technological developments are needed in all aspects of human life. It also covers the provision of data or information because it is related to someone's life. Data or information that provided in image formats is developed in health field of a human's life because it contains a lot of information about patient health that can't be described by text or audio format. Currently due to the human body is susceptible to radiation rays then as long as the patient is detected by x-ray it will be subjected to a process of radiation exposure so that the x-ray image results in a decrease (degradation) of quality. There are several operations used in image processing to solve that problem. One of them is an edge detection. Due to the current edge detection is still considered not perfect because there is still Gaussian noise. So in this paper proposed method to improve image quality from quantum noise disorder by improving the method of Sobel, so can help the medical world in detecting a disease.*

Keywords—radiation exposure, Sobel method, quality degradation, edge detection, Gaussian noise.

I. INTRODUCTION

Currently technological developments are needed in all aspects of human life. It also covers the provision of data or information because it is related to someone's life.[1] Data or information can provide in various formats such as image, text, audio, and video format. Data or information that provided in image formats is developed in health field of a human's life because it contains a lot of information about patient health that can't be described by text or audio format. [2]

For example, a technology that developed in the health field is an x-ray. The description of an x-ray is the information about patient's health. The most common issue in x-ray image detection is removing soft-tissue area or bones-tissue. The soft-tissue or bones-tissue has a negative impact to the quality of x-ray image detection.[1] Currently, the human body is susceptible to radiation rays then as long as the patient is detected by x-ray it will be subjected to a process of radiation exposure so that the x-ray image results in a decrease (degradation) of quality. The degradation of image results that caused by digital data storage also called a noise.[1]

Noise can be caused by optical instrument problem in image receiver, such as photo lens which is full of dust or unsuitable image process, such as Gaussian, Salt and Pepper, and Speckle Noise. There are several operations used in image processing to solve that problem. One of them is an edge detection.[3] Due to the current edge detection is still considered not perfect because there is still Gaussian noise, so we proposed a new method of edge detection which improve image quality from quantum noise disorder by improving the method of Sobel with Gaussian Smoothing, so the improved algorithm can help the medical world in detecting a disease. The purpose of Gaussian smoothing is to manipulating image parameters. Because of that manipulated parameters the quality of x-ray images improved.

This proposed method discuss about image segmentation especially in background removal. Main goal of this paper is to eliminating the unnecessary areas of x-ray images which aren't the main region.

II. RELATED RESEARCH

Li Hua[4] proposed an improve method of Sobel algorithm based on eight directions to improve the accuracy of Sobel operator. The eight direction used to implement an edge detection on multiple directions. The result of this experiment is a new method that reduce the lost information of an image which is the image in a smooth and continuous direction. This new method show us a better accuracy for a complex images that have a texture. The function of this method is to make sure the pixel width is closer to a single pixel width.

Lin Li[5] proposed an optimization of Sobel algorithm. Sobel operator is the most trustworthy method because it gives an efficient output to the result of the image. This method use to extract the object of the image. However, Sobel operator is a very complex process and the implementation process of this algorithm is very difficult, so it optimized Sobel operator using MBD optimization workflow.

MBD workflow can automatically generate HDL code to validate FPGA design process of image

processing. MDB workflow can accelerating the design process of image processing system, so it's development time can be reduced. Because it optimized a particular solution to compare a different solutions in order to obtain the best tradeoff between cost and performance.

Shi Tian[6] proposed a parallel computer system, to detecting surface defects of cameras by adding six different directions (45°, 135°, 180°, 225°, 270° and 315°), to ensure the grayscale output in the allowable error. It allows an accurate and efficient positioning to reduce noises, so it's extracted more efficient features and the parameters characteristic. With neural network training it defects classification and obtain precision with 106 iterative steps and time of 3 s less than the Sobel algorithm.

III. PROPOSED WORK

Image processing can be described as a processing of a signal that have an image input. The input image of this process produce a characteristic or parameter that represent the image or may be present a new image. This new image related to digital form that has been converted with image enhancement process, such as restoration, compression, etc.[3]

Image enhancement can be described as an accentuation of some image features. These features include boundaries, contrast, gray level & contrast manipulation, noise reduction, edge detection, sharpening, filtering, etc. That process doesn't increase content of this data just make this image more useful. [7]

Image restoration can be described as an image filtering to defects the effects of degradations. This process depends on the filter design of degradation process to adjust the effectiveness of this process. Image restoration and image enhancement has a different in extraction and accentuation for some image features. Image restoration has more concerned of image extraction than image enhancement. [7]

Image compression can be described as a minimizing bit process of the represented image.

On this paper we proposed a new model of edge detection which is a fundamental subject in image processing. It also included in machine vision and computer vision subject. [7]

Edge detection can be used to extract a features in some area. This algorithm (Canny, Sobel, Robert) can be used to extract some features of an image to detect some of different types of sharp changing points. In image processing the changes brightness abruptly can be defined as edges points. [7]

Edges points can be created by shadows, textures, geometry, etc. It also defined as a discontinuities of image intensity in an image structure. Edge detection can be described into four steps, such as:[7]

1. Image Smoothing process: a suppress process of a noise without defect the true images.
2. Image Enhancement: A process that applying a filter into an images to enhance the quality of image edge.
3. Image Detection: A process to determine the pixel of an edge, which is should be discarded as a noise or retained as an image. This process also include the thresholding process.
4. Image Localization: A process that determine the exact location of an edge

Due to the current edge detection is still considered not perfect because there is still Gaussian noise, so we proposed a new method of edge detection which improve image quality from quantum noise disorder by improving the method of Sobel with Gaussian Smoothing, so the improved algorithm can help the medical world in detecting a disease. The purpose of Gaussian smoothing is to manipulating image parameters. Because of that manipulated parameters the quality of x-ray images improved. [3]

This proposed method discuss about image segmentation especially in background removal. Main goal of this paper is to eliminating the unnecessary areas of x-ray images which aren't the main region.

In this research we need a research flow that will be used illustrated by the following diagram :

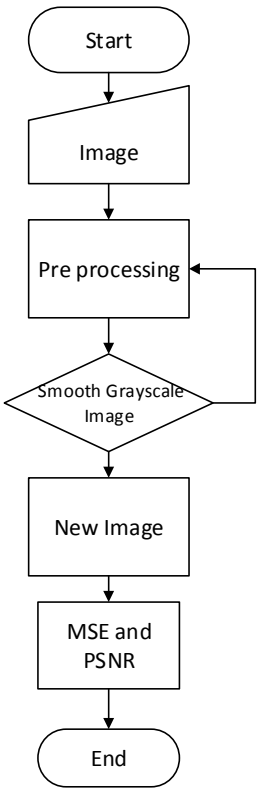


Fig. 1 Methodology Research

First process is the preprocessing that change RGB image to Grayscale which will be followed by smoothing method using Gaussian method. After that edge detection will be done by using Sobel method to produce a new image whose quality can be measured by MSE and PSNR.

A. GRAYSCALE

The image that we use for research is a grayscale image. The definition of grayscale is an image that has only one channel value, which is marked by the similarity of Red and Blue pixel values. In the initial process there should be a definition of objects and variables that will be used to accommodate the value of RGB images used. Then do the calculations horizontally and vertically each pixel and then save the value. Calculate the pixel value with the formula: 0.299 * Red, 0.587 * Green, and 0.114 * Blue. Show Grayscale results into the variables we have created at the beginning. [8]

This process can be described by this pseudocode:

Input : pixel value x in frame X = 0,..,Last Frame
Output : Image with Binary mask Value

1. Initialize model C for pixel P0 and store it in Array
2. For X = 0, Y = 0 , Last Frame
3. Get pixel RGB
4. Grayscale = (R+G+B)/3
5. Set pixel grayscale

Fig. 2 Pseudocode Process

Here is a Grayscale process diagram:

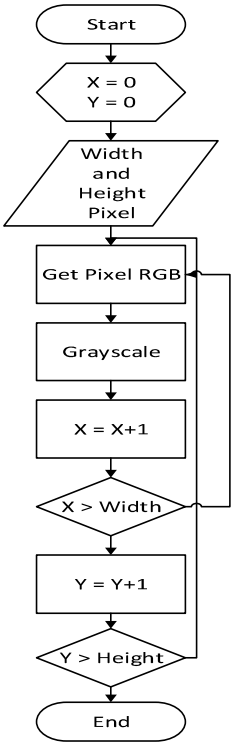


Fig. 3 Grayscale Method

B. Gaussian Smothing

Gaussian smoothing is a method that uses gaussian function and serves to smooth the noise on the image. Another mask that is often used for image refinement is the gaussian smoothing mask. The weights on the gaussian refining mask follow the normal distribution as \ which is expressed in the equation below:[9]

$$f_0(x,y) = \frac{f^r(x,y) + f^g(x,y) + f^b(x,y)}{3} \ldots\ldots(1)$$

In the Gaussian Smoothing Process the original image will be filtered using a mask with the formula:[9]

$$\frac{1}{2\pi\sigma^2} e^{\frac{-x^2+y^2}{2\sigma^2}} \ldots\ldots(2)$$

Where :

π = radius (3,14)

σ = the length of matrix

(x, y)= X and Y point pixel

e = Natural logarithm (2,7182)

After that proceed with the creation of objects and definitions of variables which will then be initialized with an array of lists that will be done scanning process horizontal and vertical scan the results will be stored in the variables that we define at the beginning of the process.[9]

Here is a Gaussian Smoothing process diagram:

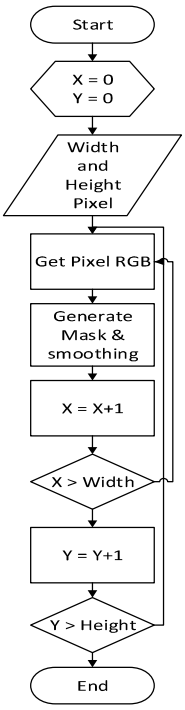


Fig. 4 Gaussian Smoothing Process

Here is a Gaussian Smoothing process pseudocode :

Input : pixel value x in frame X = 0,..,Last Frame
Output : Image with Binary mask Value
1. Initialize model C for pixel P0 and store it in Array
2. For X = 0, Y = 0 , Last Frame
3. Get pixel RGB
4. Generate Mask
5. Gaussian Smoothing
6. Set pixel Gaussian Smoothing
d

Fig. 5 Pseudocode Gaussian Smoothing Process

C. Edge Detection

Edge is a sudden change in the intensity of a sudden large degree of gray within a short distance. So the definition of edge detection is the detection of the value of the intensity change in an image. There are three kinds of edges in the digital image, namely: Steep Edge, Landed Edge and also the Edge with Noise.[10]

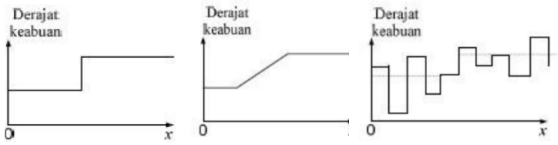


Fig. 6. Various edges a. Steep Edges b. Landed Edge c. Edges with noise

Inside this edge detection there are many different methods used in detecting the edges of this image. One of the best methods is Sobel. In edge detection, generally the image edges are considered discontinuous whose derivative values are taken at their maximum values to determine edges. Sobel edge detection is used to calculate the intensity gradient of the image and to find the possible direction of improvement from the phase of light to the dark phase and the rate of change of direction. The results show sudden or subtle changes in point intensity.[5]

In the Sobel edge detection process the first thing to do is create the object and define the variables, then initialize the Array list and also do the scanline horizontally and vertically which will then be filtered, after which the result will be displayed.[7]

Here is a pseudocode example of the Sobel edge detection process:

Input : pixel value x in frame X = 0,..,Last Frame
Output : Image with Binary mask Value
1. Initialize model C for pixel P0 and store it in Array
2. For X = 0, Y = 0 , Last Frame
3. Clear List
4. Get Neighbors List
5. Get Sobel Value
6. Threshold
7. Set pixel Sobel
8. End for

Fig. 7. Various edges a. Steep Edges b. Landed Edge c. Edges with noise

After scanning the data horizontally and vertically in the detection of the edge of a Sobel, the need for a process of combining existing data on horizontal and vertical arrays, the process is better known as Gradient Magnitude.

Here is the pseudocode of the process:

Input : pixel value x in frame X = 0,..,Last Frame
Output : Image with Binary mask Value
1. Initialize model C for pixel P0 and store it in Array
2. For X = 0, Y = 0 , Last Frame
3. Clear List
4. Get Sobel Value
5. Gradient Magnitude
6. Threshold
7. Set pixel Magnitude
8. End for

Fig. 8. Gradient Magnitude Pseudocode

IV Evaluation

A. Data Set

In this paper we used an X-ray image from public chest radiograph database of Radiological Technology (JSRT) [11].This database included 247 x-ray image that devided into two cluster. The first cluster contain 154 images and the second one contains 93 images. That images have standard size 2048 x 2048 pixel. In JSRT, images that have been taken in raw format, so if we want to open in our desired format, we must have an additional image file viewer.



Fig. 9. Original Images

B. Parameter

In image processing, quality is highly influential in the interpretation of images. In image quality processing is measured by Peak Signal to Noise Ratio (PSNR). The definition of PSNR is the ratio between the maximum value of the pixel as measured by the amount of noise that affects the pixel. PSNR is measured in decibels. PSNR can also be used to find out the quality comparison of cover image before and after the message inserted. To determine PSNR, MSE should first be determined (Mean Square Error). MSE is the mean square error value between the cover image and image, can be mathematically formulated as follows:[12]

$$MSE = \frac{1}{MN}\sum_{y=1}^{M}\sum_{x=1}^{N}[I(x,y) - I'(x,y)]^2$$

Where :
MSE = Mean Square Error
M = Length Matrix
N = Width Matrix
I(x, y) = Pixel I
Having obtained the value of MSE then the value of PSNR can be calculated from the square of the maximum value divided by MSE. Mathematically, the PSNR values are described as follows:

$$PSNR = 10.log\left(\frac{MAXi^2}{MSE}\right)$$

Where:
MSE = MSE value
MAXi = the maximum value of the image pixels that are
The lower the value of MSE, the better, and the greater the value of PSNR the better the image quality.

Here is the pseudocode of the process:

Input : pixel value x in frame X = 0,..,Last Frame
Output : Image with Binary mask Value

1. Initialize model C for pixel P0 and store it in Array
2. For X = 0, Y = 0 , Last Frame
3. Get Pixel Before and After
4. Get Sobel Value
5. Get MSE
6. Get RMSE
7. Get PSNR

Fig. 9. MSE and PSNR Pseudocode

C. Evaluation test

1. Fidelity test

Fidelity test is a test performed to determine the ability of edge detection in the parent image, which is done by looking at the difference between the initial image of the parent and the image of the parent that has done edge detection.

Observations are done visually and quantitatively. As a visual benchmark in this study is the image of the parent does not experience significant changes when viewed using the human eye by naked eye. This study aims to find out the implementation of edge detection on true color image with operator method of Sobel. By allowing the current edge detection can find the image and its classification in the future. The steps per step are as follows:

1. The selected image will be inserted into the program
2. After the image is inserted then the image is changed into grayscale



Fig. 10 Grayscale Image

3. After the image changed into grayscale form we perfomed edge detection to the image

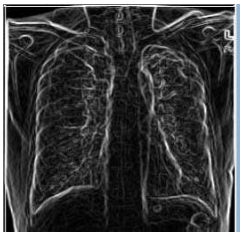


Fig. 12 Edge detection Process

4. After the image changed into edge detection form we perform Gaussian smoothing to the image

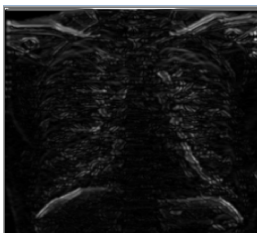


Fig. 13 Gaussian Smoothing Process

5. MSE and PSNR

In this paper the quality of produced image will be tested for durability by robustness testing as measured by various image manipulations. In this study, tests were performed with the addition of brightness and contrast.

The addition of brightness and contrast in the image, used several samples with different levels of brightness (5-20). After the added level of brightness then the image in the count MSE and PSNR it.

Table 1 MSE and PSNR for Edge detection and Edge detection with Gaussian Smoothing

Smoothing	Edge Detection	Improvement
MSE	0,431	0,1979
PSNR	119,252	127,0227

Table 2 MSE and PSNR for Edge detection with several samples of brightness

	Edge Detection		Improvement	
Nilai Contrast	MSE	PSNR	MSE	PSNR
Contrast +5	0,431	127,680	0,431	127,680
Contrast +10	0,431	127,680	0,431	127,680
Contrast +15	0,431	127,680	0,431	127,680
Contrast +20	0,431	127,680	0,431	127,680

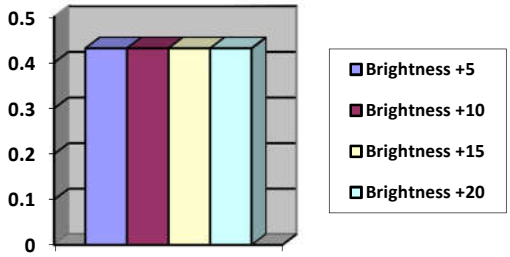


Fig. 14 MSE in different Brightness before Smoothing Process

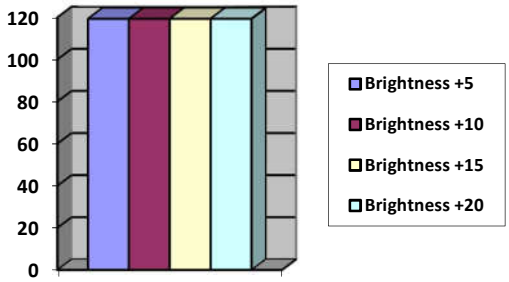


Fig. 15 PSNR in different Brightness before Smothing Process

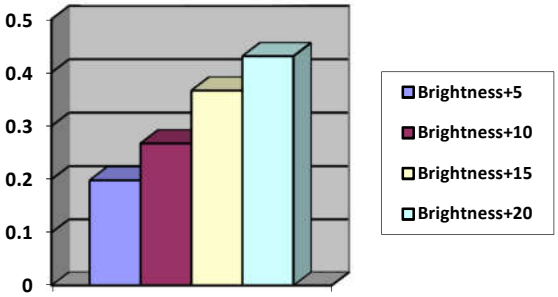


Fig. 16 MSE in different Brightness with Smothing Process

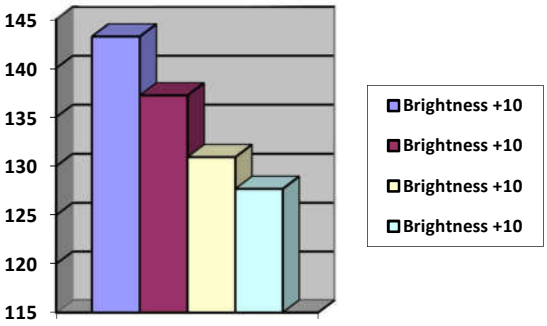


Fig. 17 PSNR in different Brightness with Smothing Process

Table 3 MSE and PSNR for Edge detection with several samples of contrast

	Edge Detection		Improvement	
Nilai Contrast	MSE	PSNR	MSE	PSNR
Contrast +5	0,431	127,680	0,1980	143,220
Contrast +10	0,431	127,680	0,2672	137,218
Contrast +15	0,431	127,680	0,3662	130,917
Contrast +20	0,431	127,680	0,4305	127,680

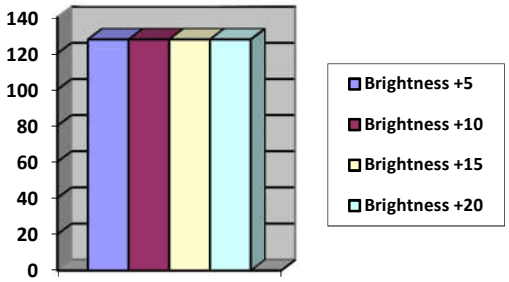


Fig. 18 MSE in different Contrast with and without Smothing Process

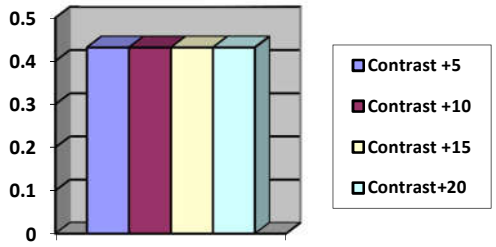


Fig. 19 PSNR in different Contrast with and without Smothing Process.

2. Robustness Test

Robustness test is the best test for edge detection. It used to compared the edge detection method and our proposed method with 20 samples images to determine which method that the best quality of MSE and PSNR. From this table we know that Gaussian Method is better because it's MSE and PSNR better than edge detection method.

From the visual test, the main image and the result image shows that the bigger the PSNR, the more the parent image and the clearer the result image.

From the test of brightness, we can conclude that the higher the brightness MSE is getting bigger and its PSNR is getting smaller

Table 4 MSE and PSNR for Edge detection with several X-ray dataset

	Edge Detection		Improvement	
	MSE	PSNR	MSE	PSNR
Image 1	0,431	119,252	0,431	119,252
Image 2	0,431	119,252	0,1979	127,0227
Image 3	0,431	119,252	0,1979	127,0227
Image 4	0,431	119,252	0,3662	120,8709
Image 5	0,431	119,252	0,431	127,680
Image 6	0,431	119,252	0,431	119,252
Image 7	0,431	119,252	0,431	119,252
Image 8	0,431	119,252	0,431	119,252
Image 9	0,431	119,252	0,431	119,252
Image 10	0	infinity	0	infinity
Image 11	0	infinity	0	infinity
Image 12	0,0049	163,912	0,0049	163,912
Image 13	1,6433	105,858	0,4475	118,866
Image 14	1,2570	108,538	0,8215	112,792
Image 15	0,5543	116,727	0,1633	128,946
Image 16	0,00495	163,912	0,00495	163,912
Image 17	0,14846	129,900	0,04950	140,8857
Image 18	1,25697	108,538	0,44539	118,9134
Image 19	1,25697	108,538	0,79674	113,0975
Image 20	0,55426	116,727	0,16331	128,9465
Average	0,528006	109,7959	0,312255	114,4564

IX. CONCLUSION

In this research, some conclusions are drawn:

a. From the visual test, the main image and the image of the result shows that the larger the MSE the more the parent image and the resulting image becomes less clear (blur).

b. From the visual test, the main image and the result image shows that the bigger the PSNR, the more the parent image and the clearer the result image.

c. From the test of brightness, we can conclude that the higher the brightness MSE is getting bigger and its PSNR is getting smaller
d. From contrast testing, we can conclude that the higher the contrast then the MSE and PSNR not change.

X. ADVICE

The suggestions that the author can provide in this study are as follows:
The uncertain results of MSE and PSNR are due to the multitude of ways MSE and PSNR search
Display results still look rough so there needs to be further filtering
The pixel processing still chooses pixels in sequence so it takes a long time so it needs to be improved in terms of accuracy and efficiency.

REFERENSI

[1] J. Romanowski and T. Nowak, "Improved X-ray Edge Detection," pp. 309–310, 2013.

[2] L. Xiao, C. Li, Z. Wu, and T. Wang, "An enhancement method for X-ray image via fuzzy noise removal and homomorphic filtering," *Neurocomputing*, vol. 195, pp. 56–64, 2016.

[3] F. S. Khan and H. Bhaskar, "Advances in the Quantum Theoretical Approach to Image r r r r r," vol. 49, no. 4, 2017.

[4] "An Edge Detection Method Based on Improved Sobel Operator Lihua Sun , Enliang Zhao, Long Ma , Li Zheng," vol. 973, pp. 1529–1532, 2014.

[5] L. Li, "An Improved Image Processing Algorithms of Sobel Edge Detection and Functional Verification," vol. 55, no. 8, pp. 359–366, 2017.

[6] T. Shi, J. yi Kong, X. dong Wang, Z. Liu, and G. Zheng, "Improved Sobel algorithm for defect detection of rail surfaces with enhanced efficiency and accuracy," *J. Cent. South Univ.*, vol. 23, no. 11, pp. 2867–2875, 2016.

[7] P. Pawar and D. Rai, "Improved Sobel Algorithm Based Image Edge Detection using Gaussian Filter," vol. 3, no. 8, pp. 561–567, 2017.

[8] K. Padmavathi and K. Thangadurai, "Implementation of RGB and grayscale images in plant leaves disease detection - Comparative study," *Indian J. Sci. Technol.*, vol. 9, no. 6, pp. 4–9, 2016.

[9] C. Pérez-Benito, S. Morillas, C. Jordán, and J. A. Conejero, "A model based on local graphs for colour images and its application for Gaussian noise smoothing," *J. Comput. Appl. Math.*, vol. 330, pp. 955–964, 2018.

[10] N. Mathur, S. Mathur, and D. Mathur, "A Novel Approach to Improve Sobel Edge Detector," *Procedia Comput. Sci.*, vol. 93, no. September, pp. 431–438, 2016.

[11] M. N. Saad and H. A. Hamid, "Image Segmentation for Lung Region in Chest X-ray Images using Edge Detection and Morphology," *4th IEEE Int. Conf. Control Syst. Comput. Eng.*, no. November, pp. 28–30, 2014.

[12] M. A. Soeleman, M. Hariadi, and M. H. Purnomo, "Adaptive threshold for background subtraction in moving object detection using Fuzzy C-Means clustering," *TENCON 2012 IEEE Reg. 10 Conf.*, pp. 1–5, 2012.

Organized by

LPPMUDINUS

Lembaga Penelitian dan Pengabdian Masyarakat Universitas Dian Nuswantoro Semarang

Computer Science Faculty
Engineering Faculty
Health Faculty
Universitas Dian Nuswantoro Semarang