International Journal of Computer Applications Technology and Research
Volume 9–Issue 02, 41-46, 2020, ISSN:-2319–8656

Forecasting of Arabica Coffee Production in Bali Province Using Support Vector Regression

Ni Made Ratna Putri Udiani
Department of Information Technology
Udayana University
Jimbaran, Indonesia

I Ketut Gede Darma Putra
Department of Information Technology
Udayana University
Jimbaran, Indonesia

Gusti Made Arya Sasmita
Department of Information Technology
Udayana University
Jimbaran, Indonesia

Abstract: Coffee is one from 40 the leading commodities of national commodities and one of superior commodities in the province of Bali. Bali Province's records shows an increasing growth from Arabica coffee plantations in the period covering 8,205 hectares in 2009 increased to 13,155 hectares in 2014 (Bali Central Bureau of Statistics, 2015). Based on the amount of Arabica coffee production that continues to increase, we need a study to find out the achievement of results and policies that will be carried out in order to increase the results of Arabica coffee production. Research on forecasting Arabica coffee production in Bali using Support Vector Regression and several kernels. Pearson Universal Kernel and RBF kernel. Forecasting in the future three years from 2019-2021 has increased. The test MAPE results using the Universal Person Kernel is 5.14% and the RBF kernel is 7.68%.

Keywords: Forecasting, SVR, PUK, RBF, Arabica Coffee

1. INTRODUCTION

Indonesia is an Agricultural Country, this can be seen from the large land area used for agriculture. Based on the existing land area in Indonesia, around 74.68% is used for agricultural land. Various research results, concluded that the biggest contribution in reducing the number of poverty is the growth of the agricultural sector. Whereas the contribution of the agricultural sector in reducing the number of poverty reached 66 percent, with about 74 percent in rural areas and 55 percent in urban areas [1].

The plantation sector has a significant potential role in the natural resource [2]. Coffee is one of the leading commodities from 40 national commodities and one of the most superior commodities in the Bali Province. Based on the Bali Province's statistics shows an increasing growth of Arabica coffee plantation's field in the period covering 8,205 hectares in 2009 increased to 13,155 hectares in 2014 (Bali Central Bureau of Statistics, 2015). With the increase in plantation area, the amount of Arabica coffee production also increased in 2009 which amounted to 3,135,750 tons and increased to 4,183,924 tons in 2014 [3].

Based on the amount of Arabica coffee production that continues to increase, research is needed to obtain results and policies that will be carried out in order to increase the results of Arabica coffee production. Research can be done by doing forecasting to find out the amount of coffee production in the future.

Several studies on Coffee Production forecasting have been carried out previously, namely research on forecasting for robusta coffee production demand at PT. XYZ. From the results of forecasting, the robusta coffee demand pattern is obtained which tends to be constant and based on forecasting calculations with the 3 Center Moving Average method, the result of forecasting the number of requests for robusta coffee is 8754MT [4]. Other studies regarding rainfall forecasting use Support Vector Regression. The results of the research evaluation conducted on the data show that the projected technique performs higher than the conventional framework in terms of accuracy. and the time process runs. The proposed approach produces a maximum prediction of 99.92% [5].

Other studies regarding forecasting the availability of food use Support Vector Regression shows three defining attributes used in this study are (1) Harvest Area; (2) Total Harvest Production; and (3) Food

From the results of experiments, it is known that the biggest contributor to food products is Java Island, especially in East Java. Almost every type of crops, East Java plays an important role in the production of food needed in Indonesia [6].

This study shows a forecasting of the amount of Arabica coffee production using a different method from the literature study described earlier in order to produce a forecast that is close to the actual data. This study uses the Support Vector Regression (SVR) method.

2. METHODOLOGY

The methodology used on this research will be explained by the sequence of research conducted. They are data collection, data normalization, forecasting by sharing training data and test data, forecasting results, data visualization, forecasting accuracy. A general chart of the research can be seen in Figure 1.

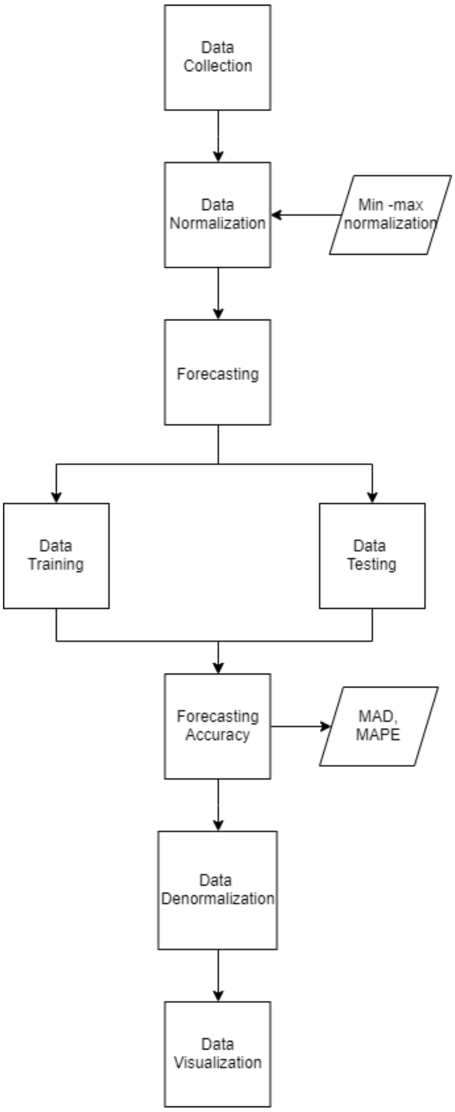


Figure 1. General Chart

Figure 1. is a general chart of research. The study began with data collection, data normalization, forecasting (test data and training data), forecasting results, data denormalization, and data visualization..

2.1 Data Collection

The Data used on this study obtained from the Bali Central Statistics Agency which can be accessed via the web address https://bali.bps.go.id/. The research data used were from 2000 to 2018 for 19 years.

2.2 Data Normalization

Normalization of data is part of data transformation, which is a technique to convert data into values that are more easily understood [7]. Normalization is very important in handling real-time data, because real-time data has different ranges and units. Normalization is used to scale real-time data in the range of 0 to 1. The normalization process helps make numerical calculations more precise and improves the accuracy of forecast results. The formula for data normalization is in Figure 2.

$$v' = \frac{v - \min_A}{\max_A - \min_A}(new\_\max_A - new\_\min_A) + new\_\min_A$$

Figure 2. Normalization Formula

V is the actual data, minA is the lowest actual data, maxA is the highest actual data, new_maxA is the highest data scale that is 1, and new_minA is the lowest data scale that is 0.

2.3 Forecasting

The forecasting will use the Weka v3.9.3. The data used in the forecasting is the amount of Arabica coffee production from 19 years (2000-2018). Forecasting is done by dividing the data into two parts, they are training data and test data. Training data used in 2000-2012. Test data used in 2013-2018. Data predicted in 2019-2021 are measured for accuracy using MAPE percentages.

This study uses the Support Vector Regression method. The Support Vector Regression method in this study uses two kernels, namely Pearson Universal Kernel, and RBF Kernel

2.4 Forecasting Accuracy

The accuracy of forecasting is shown by the percentage of MAPE. Mean Absolute Percentage Error (MAPE) is the average absolute percent error, where the absolute deviation value for each period is divided by the actual value for that period. The formula for calculating MAPE is as follows.

$$\text{MAPE} = \frac{\sum((\text{deviation absolute}/\text{actual data})\text{x}100)}{n}$$

MAPE (Mean Absolute Percentage Error) is the easiest step to interpret. For example, the MAPE 2% result is a clear statement that does not depend on problems such as the size of the input data. According to Lewis, C.D. (1982), the interpretation of MAPE is divided into 4 categories [8], as shown in Table 1.

Table 1. MAPE Intepretation

MAPE Value	Interpretation
< 10%	Very Good
10 - 20%	Good
20 - 50%	Enough
>50%	Bad

Table 1 is an interpretation of MAPE. The MAPE category consists of four divisions: very good, good, enough, and bad.

2.5 Data Denormalization

Forecasting results at WEKA are data with a scale of 0-1. Data must be normalized so that it can be compared with actual research data. The equation used in min-max denormalization is as follows.

Denormalization = ($normalizedValue * ($max-$min) + $min)

Where $normalizedValue is the normalized value, $max is the highest actual data, and $min is the lowest actual data.

2.6 Data Visualization

Data visualization is important to do in presenting data. Data visualization is useful to facilitate understanding of data, as well as being communicative and more interesting. This research visualizes data in graphical form with the help of Microsoft Excel.

3. THEORY AND CONCEPTS

Concepts and theories contain reference material that is used according to the topic of research conducted. The topic of the research was forecasting Arabica coffee production. References that are loaded are forecasting, SVR, and the kernel.

3.1 Forecasting

Forecasting is an important issue that covers many fields including business and industry, government, economics, environmental science, medicine, social sciences, politics, and finance. Forecasting problems are often classified as short term, medium term, and long term. Short-term forecasting problems involve predictive events for only a few time periods (days, weeks, months) in the future. The medium-term forecast extends from one to two years into the future, and the problem of long-term forecasting can extend beyond the next few years. Short and medium term forecasting is needed for activities that range from operations management to budgeting and selection of new research and development projects. Long-term forecasting impacts on issues such as strategic planning. Short and medium term forecasting is usually based on the identification, modeling and extrapolation of patterns found in historical data. This historical data usually shows inertia and does not change dramatically very quickly [9].

3.2 Support Vector Regression

Support Vector Regression (SVR) is the improvement from time series forecasting and forecasting in various domains, including business and management science [10]. Support Vector Regression is a Support Vector Machine method for regression cases. The basic problem of regression is to determine a function that can accurately predict future values. This can be done by forming the dividing plane with the smallest size while minimizing the amount of distance between data points to the separating plane [11].

Support Vector Regression has parameter C. Parameter C is the penalty value for SVR model errors.

3.3 Kernel

The kernel method is a class of algorithms that developed in the 90s in the field of machine learning. The best-known example of the kernel method is Support Vector Machines (SVMs), which are good for classification problems. The kernels used in this study are Pearson Universal Kernel and RBF [12].

3.3.1 Pearson Universal Kernel

Pearson Universal Kernel is famous in the field of spectroscopy. The Pearson Universal Kernel function was taken as an alternative kernel function in this study. Pearson Universal Kernel functions as a kind of universal kernel that can replace (by selecting the appropriate parameter settings) kernel functions that are commonly applied, namely the Linear, Polynomial, Gaussian and Sigmoid kernels. The function of the Pearson Universal Kernel is as follows.

$$K(x_i, x_j) = 1/[1+(2\sqrt{\left\|x_i - x_j\right\|^2}\sqrt{2^{(1/\omega)}-1}/\sigma)^2]^\omega$$

xi and xj are two vectors. The Pearson VII kernel function will point to the symmetrical matrix with the diagonal and all other entries ranging between 0 and 1 for each random pair (xi and xj). PUK has the resistance and strength in mapping the same or even stronger compared to standard kernel functions, which results in the same or better generalization performance than SVM [13].

3.3.2 RBF

The RBF kernel is a kernel that can generally be used for all types of data. This kernel uses the RBF kernel function to get inside x and y using the equation.

K(x, y) = exp(-gamma*(x-y)^2)

where x, y are data values in two different feature spaces and σ (sigma) is a free parameter in the RBF kernel that determines kernel weight. The parameter σ needs to be adjusted to provide more accurate classification results. The default value σ is 1. The use of the gamma parameter can be used on the RBF function whose value $\gamma = \frac{1}{\sigma^2}$ [14].

4. RESULT AND DISCCUSION

Forecasting the case of the amount of Arabica coffee production with several kernels. Arabica coffee production data in Bali is 19 years. Forecasting results are checked for accuracy and compared with actual data using graphs.

4.1 Pearson Universal Kernel

Forecasting uses Arabica coffee production data in Bali using the Support Vector Regression and Pearson Universal Kernel methods. Forecasting Arabica coffee production in Bali is done by finding the best parameters to produce the lowest percentage of MAPE. The results of testing the parameters used are in Table 2.

Table 2. C Testing

C	Error Percentage (MAPE%)
1	5.14%
10	15.30%
100	15.30%

Table 2 is the result of a comparison of several parameter tests. The parameters used in the Support Vector Regression and Pearson Universal Kernel methods are C. Magnitude C affects the test results. C is smaller, resulting in a better percentage of MAPE. The best percentage of MAPE produced was 5.14%, with C 1.0. The PUK kernel has parameters namely sigma and omega. The best sigma and omega test results are in Table 3.

Table 3. PUK Parameter Testing Result

Sigma	Omega	Error Percentage (MAPE%)
0.1	0.1	5.14%
1.0	1.0	13.69%
2.0	2.0	17.21%
3.0	3.0	20.97%

Table 3 is the result of a comparison of several parameter tests. The parameters used in the Support Vector Regression and Pearson Universal Kernel methods are sigma and omega. The amount of sigma and omega affects the test results. Sigma and omega are smaller, resulting in a better percentage of MAPE. The best percentage of MAPE produced was 5.14%, with sigma 0.1 and omega 0.1. Forecasting results on Arabica coffee production are in Table 4.

Tabel 4. Arabica Coffee Forecasting Result

Year	Actual	Training	Testing	MAD	MAPE
2000	6796				
2001	5644	5639.946			
2002	3768	3774.945			
2003	4413	4406.493			
2004	3696	3700.428			
2005	3279	3280.905			
2006	2679	2682.294			
2007	3296	3304.372			
2008	3136	3140.927			
2009	3475	3481.403			
2010	3485	3485.109			
2011	3123	3127.341			
2012	4200	4195.291			
2013	4215		3897.632	317.368	7.53%
2014	3804		3928.51	124.5095	3.27%
2015	4154		3944.566	209.4342	5.04%
2016	4052		3967.209	84.7907	2.09%
2017	3473		3733.364	260.3637	7.50%
2018	4217		3988.206	228.794	5.43%
2019			4003.851		
2020			4093.19		
2021			4141.77		
Average				**204.21**	**5.14%**

Table 4 is the result of forecasting Arabica coffee production data in Bali with Support Vector Regression and Pearson Universal Kernel. The results of the performance of Arabica coffee production using Support Vector Regression and RBF kernels are "very good" because they are below 10%, according to Lewis, C.D. (1982) [8]. Visualization of forecast results is illustrated using a line graph that can be seen in Figure 3.

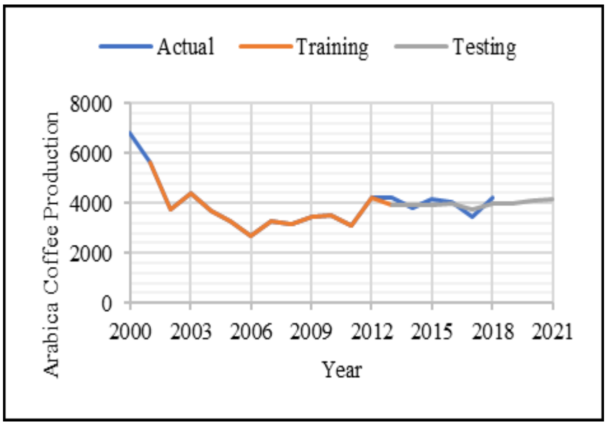


Figure 3. Forecasting Result With SVR and PUK

Figure 3 is a graph of forecasting results on the amount of Arabica coffee production using Support Vector Regression and Pearson Universal Kernel Forecasting results consisting of three years (2019-2021) which have increased. Forecasting using Support Vector Regression and Pearson Universal Kernel 1 is influenced by the magnitude of the parameters C, sigma and omega which can be seen in Table 2 and Table 3

4.2 RBF Kernel

Forecasting uses Arabica coffee production data in Bali using the Support Vector Regression and RBF Kernel methods. Forecasting Arabica coffee production in Bali is done by finding the best parameters to produce the lowest percentage of MAPE. The results of testing the parameters used are in Table 5.

Table 5. C Testing

C	Error Percentage (MAPE%)
1	7.68%
10	33.25%
100	52.52%

Table 5 is the result of a comparison of several parameter tests. The parameters used in the Support Vector Regression and RBF Kernel methods are C. Magnitude C affects the test results. C is smaller, resulting in a better percentage of MAPE. The best percentage of MAPE produced was 7.68%, with C 1.0. The PUK kernel has parameters namely sigma and omega. The best sigma and omega test results are in Table 5. The RBF kernel has the parameters gamma. The best gamma test results are in Table 5.

Table 6. PUK Parameter Testing Result

Gamma	Error Percentage (MAPE%)
0.1	7.68%
1.0	16.59%
2.0	16.82%
3.0	16.29%

Table 6 is the result of comparison testing of several parameters. The parameters used in the Support Vector Regression and RBF Kernel methods are gamma. The amount of gamma affects the test results. Smaller gamma, results in a better percentage of MAPE. The best percentage of MAPE produced was 7.68%, with gamma 0.1. Forecasting results on Arabica coffee production are in Table 7.

Table 7. Arabica Coffee Forecasting Result

Year	Actual	Training	Testing	MAD	MAPE
2000	6796				
2001	5644	4355.442			
2002	3768	4068.488			
2003	4413	3665.022			
2004	3696	3700.016			
2005	3279	3522.985			
2006	2679	3000.126			
2007	3296	3300.255			
2008	3136	3339.367			
2009	3475	3298.609			
2010	3485	3325.369			
2011	3123	3446.821			
2012	4200	4117.48			
2013	4215		3915.747	317.368	7.53%
2014	3804		3536.571	124.5095	3.27%
2015	4154		3953.212	209.4342	5.04%
2016	4052		3648.965	84.7907	2.09%
2017	3473		3666.257	260.3637	7.50%
2018	4217		3726.777	228.794	5.43%
2019			4060.665		
2020			4077.957		
2021			4138.477		
Total				**308.9976**	**7.68%**

Table 7 is the result of forecasting Arabica coffee production data in Bali with Support Vector Regression and RBF Kernel. The results of the performance of Arabica coffee production using Support Vector Regression and RBF kernels are "very good" because they are below 10%, according to Lewis, C.D. (1982) [8]. Visualization of the results of the forecasting is illustrated using a line graph that can be seen in Figure 4.

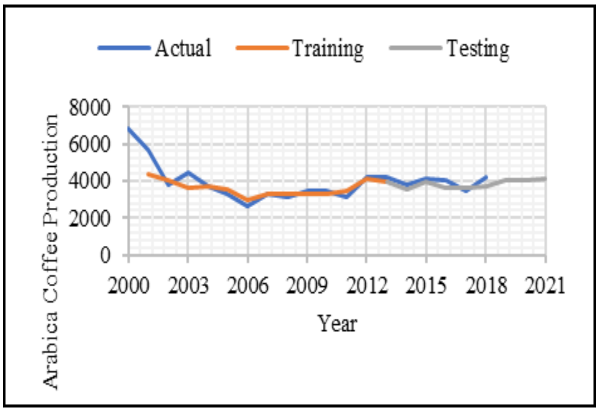


Figure 4. SVR and PUK Forecasting Result

Figure 4 is a graph of forecasting results on the amount of Arabica coffee production using Support Vector Regression and RBF Kernel. Forecasting results consisted of three years (2019-2021) which experienced an increase. Forecasting using Support Vector Regression and RBF Kernel is influenced by the magnitude of the parameters C, sigma and gamma which can be seen in Table 5.

5. CONCLUSION

Coffee is one of the leading commodities of 40 national commodities in the province of Bali. Statistics on the Province of Bali show an increase in the area of Arabica coffee plantations in the period covering 8,205 hectares in 2009 increased to 13,155 hectares in 2014 (Bali Central Bureau of Statistics, 2015). With the increase in plantation area, the amount of Arabica coffee production also increased in 2009 which amounted to 3,135,750 tons and increased to 4,183,924 tons in 2014 [3].

Based on the amount of Arabica coffee production that continues to increase, we need a study to find out the achievement of results and policies that will be carried out in order to increase the results of Arabica coffee production. Research can be done by doing forecasting to find out the amount of coffee production in the future.

Research on forecasting Arabica coffee production in Bali using Support Vector Regression and several kernels. Pearson Universal Kernel and RBF kernel. Forecasting in the future three years from 2019-2021 has increased. The test results using the Universal Person Kernel is 5.14% and the RBF kernel is 7.68%.

6. ACKNOWLEDGMENTS

The author is very grateful for the Bali Central Statistics Agency for its support in providing data for this research. We also thank the research assistants who provided their input and suggestions in this study.

7. REFERENCES

[1] E. D. Martauli, “Analysis of Coffee Production in Indonesia Analisis Produksi Kopi Diindonesia,” vol. 01, no. 02, p. 2, 2018.

[2] R. R. Novanda *et al.*, “A Comparison of Various Forecasting Techniques for Coffee Prices,” *J. Phys. Conf. Ser.*, vol. 1114, no. 1, 2018.

[3] I. G. B. Udayana, “Marketing Strategies Arabica Coffee with Information Technology in Kintamani District Bangli,” *Int. Res. J. Eng. IT Sci. Res.*, vol. 3, no. 3, pp. 93–102, 2017.

[4] S. R. P. N. Hidayatika, “Usulan Penggunaan Metode Forecasting untuk Permintaan Kopi Robusta Pada PT. XYZ,” *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2013.

[5] N. Hasan, N. C. Nath, and R. I. Rasel, “A Support Vector Regression Model for Forecasting Rainfall,” *2nd Int. Conf. Electr. Inf. Commun. Technol. EICT 2015*, no. September 2016, pp. 554–559, 2016.

[6] S. D. Agustina, Mustakim, Okfalisa, C. Bella, and M. A. Ramadhan, “Support Vector Regression Algorithm Modeling to Predict the Availability of Foodstuff in Indonesia to Face the Demographic Bonus,” *J. Phys. Conf. Ser.*, vol. 1028, no. 1, 2018.

[7] R. Rismala, “Prediksi Harga Saham menggunakan Support Vector Regression dan Firefly Algorithm,” vol. 2, no. 2, p. 15, 2015.

[8] J. J. Montaño Moreno, A. Palmer Pol, A. Sesé Abad, and B. Cajal Blasco, “Using the R-MAPE Index as a Resistant Measure of Forecast Accuracy,” *Psicothema*, Vol. 25, No. 4, Pp. 500–506, 2013.

[9] D. C. Montgomery, C. L. Jennings, And M. Kulahci, *Introduction to Time Series Analysis and Forecasting*. 2008.

[10] S. F. Crone, J. Guajardo, and R. Weber, “A study on the ability of Support Vector Regression and Neural Networks to forecast basic time series patterns,” *IFIP Int. Fed. Inf. Process.*, vol. 217, pp. 149–158, 2006.

[11] T. B. Trafalis and H. Ince, “Support Vector Machine for Regression Applications to Financial Forecasting,” *Inst. Electr. Electron. Eng.*, 2000.

[12] G. Rubio, H. Pomares, L. J. Herrera, and I. Rojas,

"Kernel Methods Applied to Time Series Forecasting," *Comput. Ambient Intell.*, pp. 782–789, 2007.

[13] K. A. A. Abakar and C. Yu, "Performance of SVM based on PUK kernel in comparison to SVM based on RBF kernel in prediction of yarn tenacity," *Indian J. Fibre Text. Res.*, vol. 39, no. 1, pp. 55–59, 2014.

[14] R. Indraswari and A. Arifin, "RBF Kernel Optimization Method with Particle Swarm Optimization on SVM using The Analysis of Input Data's Movement," vol. 1, no. December, pp. 36–42, 2017.

The “Reminder” and “Online Booking” Features in the Android-Based Motorcycle Repair Shop Marketplace

Wayan Dony Mahendra
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

I Made Sukarsa
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

AA.Kt.Agung Cahyawan
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

Abstract: Generally, vehicle service is a must for the vehicle owner. However, due to tight work routines, people often forget to service their vehicles. In addition, the service process is still using a manual system, such as taking a queue number which leads to the long queue of the service time. An Android-based Motorcycle Repair Shop Information System provides a solution to remind people to do a regular service on their vehicles with a reminder feature and make online bookings. The system development uses the SDLC (System Development Life Cycle) method. The implementation process requires an Android smartphone and a computer device by using MySQL as data storage, Firebase as a notification sender, React native and Visual Studio Code are used for developing the system. The results of the UAT test (user acceptance testing) from 20 users show 55,8% answered agree to the display, features and flow of the system, 39,5% answered strongly agree to the three question parameters, and 4,7% answered disagree with the flow and display of the system.
.

Keywords: Information Systems, Reminders, Online Booking, Motorcycle Repair Shop, Android

1. INTRODUCTION

Indonesian people have a tight routine activities in their work and daily lives. They mostly use a motorcycle as their vehicles in their daily activities or going to work. The reason people choose motorcycle than cars are due to the heavy traffic that postpone them from getting to their destination. Along with the busy routine activities and work, most motorcycle riders often forget to do service for their vehicles which should be done routinely. They also complained about the long service queue time that makes them ignore the time for vehicle service. The factors that cause this problem are due to time constraints which disrupt work time. In addition, forgetting the schedule of vehicle service is another factor that often happened by motorcycle riders in general.

2. LITERATURE REVIEW

Various research and solutions related to online reminders and bookings have been done by Patil Apurva A., Patil Gautami R., Patil Mansi S., Patil Renuka H. It explains the application to solve the problem for a long time waiting in the restaurant. The application will display information about various restaurants that are nearby or far. This application will track the user's current location by using GPS and provide a list of the closest restaurants from their current location [1].

Other studies on Android-based applications is conducted by Deepti Ameta, Kalpana Mudaliar and Palak Patel. It implements an automatic alarm system for the patients, therefore they do not need to recall the time to take their medicines and they can set an alarm at the dose timing. The alarm can be set for several medicine and time settings, including date, time and description of the medicine [2].

A research conducted by I Gusti Made Satriya Wibawa develops an expired reminder system application with the Android platform. It is built and equipped with GIS (Geographic Information System) features to store and access locations that users need [3].

A research conducted by Mamay Syani, Nindi Werstantia explains about an android-based catering ordering application that aim to help customers in booking process of the catering without necessarily come to the location. This application is also developed for time and energy efficiency, as well as to get accurate information [4].

A research concerned with an Android-based Futsal Field Rental Application by Dwi Ratnasari, Hayatulloh Firman Hadi, and Jian Budiarto aims to help booking process and down payment directly on the application [5].

The next research is conducted by Findra Kartika Sari Dewi, Theresia Devi Indriasari, Yoris Prayogo regarding to an application of academic activity schedule reminders for lecturers and students. The application will provide a notification if there is a change in schedule on academic activities [6].

Kamaruddin Tone also conducted a research about an android-based class reminder application that aims to remind students about exam schedules, lecture schedules and assignment deadline schedules [7].

A research conducted by Ade Reza Pahlevi, Nur Ismawati, ST, M.Cs developed an android-based music application that aims to remind the preferred musician's gig schedule [8].

A research conducted by A research by Marlince N.K Nababan, Ricky Sandi Putra, Novi A.D Hutagaol discussed an Android-based hotel room booking applications that aims to help the people in ordering and obtaining information about vacant hotel rooms [9].

Another research by Fani Panca Sari discussed an Android-based chef food ordering application that aims to help users in choosing a chef and food to be cooked as well as bring it home [10].

The next research is from Mr. Swapnil S. Nate, Mr. Pravin S. Navele, Mr. Vikas B. Mote, Prof. Laxman S. Naik which explains about the features which consist of three types of reminders namely basic reminders, schedule reminders and medication intake reminders, which will later work according to the schedule specified by the user. The main purpose of this application is to allow users to create reminders based on location and will later notify them automatically [11].

A research conducted by Prof. VB Dhore, Surabhi Thakar, Prajakta Kulkarni, Rasika Thorat aims to design and implement a food ordering system remotely, where customers can order food before visiting the restaurant, in addition, ordering tables and also payments [12].

Andy Fred Wali and Len Tiu Wright conducted a research that can be used as a reference and support in implementing the reminder features and online booking features found in the Motorcycle Repair Shop application by learning how to apply CRM effectively to improve service quality in accordance with the research [13].

A research conducted by Susmitha Shree Lakshmi.S discusses an Android application that aims to establish communication between hospitals and patients. Customers can send requests to hospitals, therefore they can communicate by using the token number that has been given to the application. They also can find a list of hospitals that are nearby [14].

Based on research that has been done before, this study aims to design and build a motorcycle repair shop information system by applying the reminder and online booking features. The reminder feature aims to remind users who often forget to do service, while the online booking feature aims to make online booking services to reduce the length of vehicle service queues at a repair shop.

3. RESEARCH METHODS

Motorcycle repair shop information system uses the SDLC (System Development Life Cycle) waterfall model as a system design. The SDLC method is a process for developing information systems consisting of 5 stages. The stages in the waterfall model are Analysis, Design, Implementation, Testing, and Maintenance. The stages of the research can be seen in Figure 1. [15]

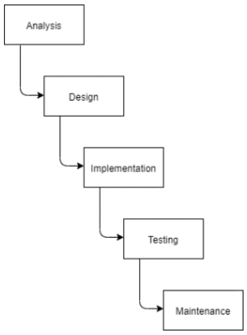


Figure 1. Research Stages

Figure 1 is the stage of system development by using the waterfall model SDLC (System Development Life Cycle) method. The analysis is the stage performed to determine the design of the application in order to answer all the needs needed by both parties, namely the repair shop and the customer. The second stage is design or UI display of the repair shop information system that uses native-based as a layout in the design or display of the system. The third stage is the creation of a motorcycle repair shop information system. The implementation process requires an Android smartphone and computer devices by using Web Service, SQLYog, Firebase, Visual Studio Code and React Native. The fourth stage is testing the developed system. It will be tested to find out the error contained in the system, and if there are many errors or malfunctions in it, a redesigning workflow will be performed to fix the system errors, therefore they can run as expected. The fifth stage is the maintenance of the system to keep the system running properly, improve the system and the performance of the developed system.

3.1 General Overview of the System

The research applications for Android-based Motorcycle Repair Shop Information Systems have a general overview that can be seen in Figure1.

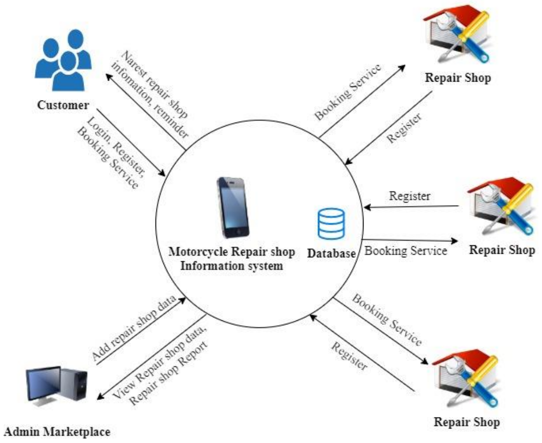


Figure. 2 General Overview of the System

Figure 2 is a general description of the system of android-based motorcycle repair shop information system. It explains the flow in making a booking service by using the application. The customers can book service through the application by registering or logging in if they already have an account. The motorcycle repair shop information system displays some of the closest repair shop locations and the customers can choose the suitable one. Furthermore, payment of the booking service is done at the repair shop chosen by the customer. The application also has a ‘reminder’ feature that will remind customers to service their vehicles regularly. It is in the form of a pop-up notification that will appear on the user's phone. The repair shop can register themselves on the application. The admin will confirm their registration and it will be displayed in the application. The data of the registered repair shop will be stored in the application database.

3.2 Context Diagram

A context diagram illustrates the whole process in the system. The scope of the Motorcycle Repair Shop Information System is presented with a context diagram. The context diagram will be explained in Figure 3.

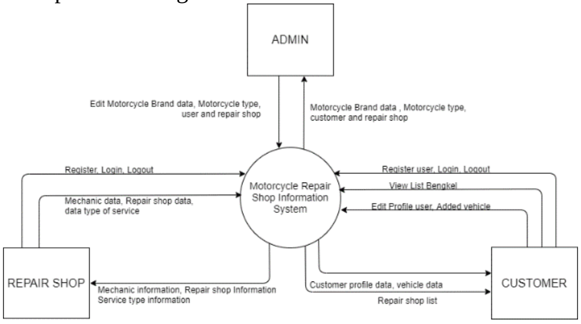

ADMIN
Edit Motorcycle Brand data, Motorcycle type, user and repair shop
Motorcycle Brand data , Motorcycle type, customer and repair shop
Register, Login, Logout
Mechanic data, Repair shop data, data type of service
Motorcycle Repair Shop Information System
Register user, Login, Logout
View List Bengkel
Edit Profile user, Added vehicle
REPAIR SHOP
Mechanic information, Repair shop Information Service type information
Customer profile data, vehicle data
Repair shop list
CUSTOMER

Figure. 3 System Context Diagram

Figure 3 is the context diagram of the Motorcycle Repair Shop Information System. The application will be used by 3 actors namely the Repair shop, the Customer and the Admin. The repair shop can use the system to register, log in to the system, log out of the system, input mechanical repair shop data, manage repair shop and service type data. In addition, they can obtain mechanic information, repair shop information and their available service type. The customer can use the system to register, log in to the system, log out of the system, view the list of repair shop, manage customer profiles, and add customer vehicles. In addition, they can get customer profile information, customer vehicle data and repair shop list. The admin can use the system to edit data of motorcycle brand, motorcycle type, customer and repair shop. In addition, they can obtain data in the form of motorcycle brand data, type of motorcycle, registered customer data and registered repair shop data in the Motorcycle Repair Shop Information System.

4. CONCEPTS AND THEORIES

This section contains concepts and theories that support in conducting the research. They are including Reminder, Android, MySQL, Firebase, React Native and Cloud computing. It will be discussed as follows.

4.1 Reminder

'Reminder' is a message feature that can help people in remembering something, it is usually found on a mobile phone or other recording media. The 'reminder' is different from the alarm that only rings at a certain time. It can be set at a certain time while displaying messages that have been written previously. The 'reminder' application can show notifications and sounds from mobile devices that aims as reminders of a schedule or agenda. In general, it is usually set by the customer based on the time when it is appeared. Notifications also can appear at certain hours or days based on the agenda entered by the customer [16].

4.2 Android

Android is a Linux-based operating system used for cellular phones (mobile), such as smartphones and tablet computers (PDAs). Android provides an open platform for developers to create their own applications that are used by various mobile devices. Since its appearance on 9th of March 2009, Android has come with version 1.1 until the last version named 5.0 Lollipop [3].

4.3 MySQL

MySQL is a popular database management system (DBMS) that has a function as a relational database management system (RDBMS). The MySQL software is an open source application. Furthermore, the MySQL database server has a very fast, reliable as well as easy to use performance and it works with client server architecture or embedded systems. It is suitable for demonstrating the database replication process due to the factor of open source and popularity.

MySQL is a database that contains one or a number of tables. The table consists of a number of rows and each row contains a table or more. The table consists of a number of rows and each row contains a table or more.

4.4 Firebase

Firebase is a back-end cloud service provider based in San Francisco, California. It makes a number of products for developing mobile or web applications. It was founded by Andrew Lee and James Tamplin in 2011 and was launched with a realtime cloud database in 2012. The main product of Firebase is a database that provides an API to enable developers to store and synchronize data through multiple clients. This company was acquired by Google in October 2014 [17].

4.5 React native

React Native is a JavaScript-based framework for creating mobile-based applications, both Android and iOS. It is a collection of JavaScript-based libraries developed by Facebook. Native React Syntax is a combination of JavaScript and XML which can be called as JSX. The React Native is a framework developed by Facebook in 2015. It was created with the aim of making it easier for web developers to create mobile-based applications, both Android and iOS. In addition, it has similarities with React for the web (ReactJS) [18].

4.6 Cloud Computing

Cloud Computing can be interpreted as a model that allows networks to be accessed easily as needed in various locations. It allows to collect computing resources, such as networks, servers, storage, applications and services in one container. According to a paper published by IEEE Internet Computing in 2008, Cloud Computing is a paradigm where information is permanently stored on a server (on the Internet) and temporarily stored in a customer's computer (client) including desktops, tablet computers, notebooks, sensors, etc [19].

5. RESULT AND DISCUSSION

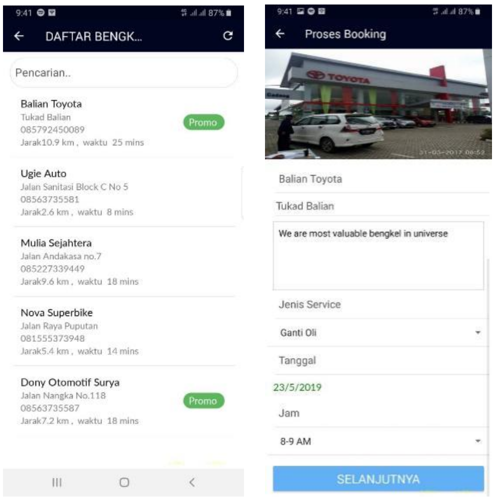
The results and discussion of the Motorcycle Repair Shop Information System application contains the results of testing the system directly, the results of Black Box testing and the results of the analysis of data development. These three results will be discussed as follows.

5.1 System Testing

The Motorcycle Repair Shop Information System has two main features, namely the online booking feature for online booking service and the reminder feature to remind the users of their regular service. These features will be explained as follows.

5.1.1 Booking Online Feature

The customer can choose the repair shop and the time in the motorcycle repair shop application. A test to make online bookings is shown in Figure 4.



Balian Toyota
Tukad Balian
085792450089
Jarak10.9 km , waktu 25 mins

Ugie Auto
Jalan Sanitasi Block C No 5
08563735581
Jarak2.6 km , waktu 8 mins

Mulia Sejahtera
Jalan Andakasa no.7
085227339449
Jarak9.6 km , waktu 18 mins

Nova Superbike
Jalan Raya Puputan
081555373948
Jarak5.4 km , waktu 14 mins

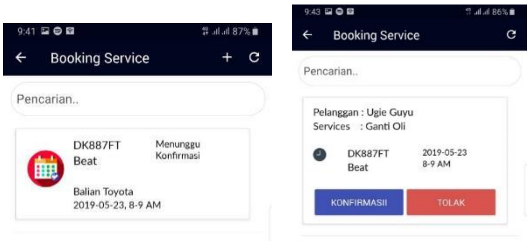
Dony Otomotif Surya
Jalan Nangka No.118
08563735587
Jarak7.2 km , waktu 18 mins

Balian Toyota
Tukad Balian
We are most valuable bengkel in universe
Jenis Service: Ganti Oli
Tanggal: 23/5/2019
Jam: 8-9 AM

(a) (b)

Figure. 4 Repair shop list and booking hours

The customer can choose the existing repair shop in the application as shown in figure 4 (a) after choosing it, they choose the type of service to be performed on their vehicles and then choose the time that is available to make a booking as shown in figure 4 (b). Confirmation of booking receipt from the repair shop will be displayed in Figure 5.



DK887FT
Beat
Menunggu Konfirmasi
Balian Toyota
2019-05-23, 8-9 AM

Pelanggan : Ugie Guyu
Services : Ganti Oli
DK887FT
Beat
2019-05-23
8-9 AM

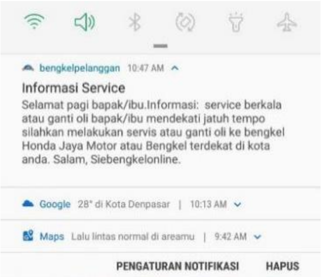
(a) (b)

Figure. 5 Customer online booking

Figure 5 (a) explains that the customer will wait for confirmation from the workshop after filling in the data to make an online booking that has been shown in Figure 4 (b). the workshop can accept or reject bookings made by the customer as shown in Figure 5 (b).

5.1.2 Reminder Feature

The reminder feature is a feature on a motorcycle repair shop application that aims to remind users of doing regular service on motor vehicles. It will be displayed in figure 6.



bengkelpelanggan 10:47 AM

Informasi Service

Selamat pagi bapak/ibu.Informasi: service berkala atau ganti oli bapak/ibu mendekati jatuh tempo silahkan melakukan servis atau ganti oli ke bengkel Honda Jaya Motor atau Bengkel terdekat di kota anda. Salam, Siebengkelonline.

Google 28° di Kota Denpasar | 10:13 AM

Maps Lalu lintas normal di areamu | 9:42 AM

Figure 6. Reminder feature

Figure 6 is a notification display that will be received by the user within the specified time after making a booking service on the information system. The users will receive a notification after the first service is done and it is set by default by the repair shop which is in the next two months. The next reminder will be received by the user based on the average service time performed by that user.

5.2 UAT (User Acceptance Testing)

User Acceptance Testing is a stage of testing the system made in a number of questions to the user to find out the deficiencies of the develop system. Question UAT (user acceptance testing) can be seen in table 1.

Table 1. UAT questions

No	Question	SS	S	TS	STS
1.	Is the form on the application (login, registration, booking service, etc.) easy to understand?				
2.	Do the icons in the application present their functions and are easy to understand?				
3.	Is the overall button layout in this application good?				
4.	Is the information displayed on the application easier to understand?				
5.	Is the color combination displayed on this application comfortable to see?				
6.	Overall, is the display of this application good?				
7.	Does the Transaction History feature in the application describe your transaction history?				
8.	Can the Online Booking feature on the application help you with your service?				
9.	Is the profile feature on the application able to describe yourself fully?				
10.	Does the Reminder feature help you as a reminder of service time?				
11.	Can the Deposit feature on the application help you make payments?				
12.	Is the My Vehicle feature in the application useful for you to find out the next service time?				
13.	Does the process for creating a new account on this application have an easy and fast process?				
14.	Does the process of booking a service (with or without promotion) take a little time and is not complicated?				
15.	Does the process for managing personal vehicle data (add, edit, delete) on the application have a complicated process?				
16.	Is the process for rating / reviewing workshops easy to do?				
17.	Does the process for adding a deposit balance to the application have an easy and fast process?				
18.	Is the application able to run responsively and there are no system errors when using the application (error, force closed, data not saved, etc.)?				

5.3 The result of UAT

The result of UAT (user acceptance testing) is based on the result of application testing towards the customer (user). The test was distributed to 20 customers (users) with 18 questions divided into 3 parameters namely display, features and system flow related to the Motorcycle Repair shop Application. The results showed 55,8% answered agree to the display, features, system flow, 39,5% answered strongly agree to the three parameter questions, and 4,7% answered disagree with the flow and display of the system.

Table 2. The result of UAT (user acceptance testing)

No	Stongly Agree (SS)	Agree (S)	Disagree (TS)	Strongly Disagree (STS)
1	9	11	0	0
2	7	12	1	0
3	6	12	2	0
4	8	12	0	0
5	6	13	1	0
6	7	10	3	0
7	5	14	1	0
8	11	9	0	0
9	8	11	1	0
10	14	6	0	0
11	7	11	2	0
12	9	10	1	0
13	6	14	0	0
14	9	11	0	0
15	7	13	0	0
16	6	14	0	0
17	9	8	3	0
18	8	10	2	0
Total UAT Percentage: 39,5% Strongly Agree, 55,8% Agree, 4,7% Disagree				

6. CONCLUSION

An Android-Based Motorcycle Repair Information System with ‘reminder’ and ‘online booking’ features are designed and applied in the form of a mobile application. The motorcycle repair shop application is designed as a media to remind service time and reduce the length of service queues. The motorcycle repair shop information system is built specifically on mobile devices with the Android platform that can be used by customers (users). It was developed using the SDLC (System Development Life Cycle) method. The process of storing and processing data the system uses MySQL to support database services. The Android-based motorcycle repair shop information system with ‘reminder’ feature utilizes Firebase as a media for sending reminder notifications to the application on Android mobile devices for customers and the system development uses React Native and Visual Studio Code. The result of UAT (user acceptance testing) from 20 users showed 55,8% answered agree to the display, features, system flow, 39,5% answered strongly agree to the three parameter questions, and 4,7% answered disagree with the flow and display of the system.

REFERENCES

[1] P. A. A *et al.*, “Location Based Restaurant Seat Booking Application For Android Phones : An overview,” 2017.

[2] D. Ameta, K. Mudaliar, and P. Patel, “Medication Reminder and Healthcare – an Android Application,” *Int. J. Manag. Public Sect. Inf. Commun. Technol.*, vol. 6, no. 2, pp. 39–48, 2015.

[3] I. G. Made, S. Wibawa, I. M. Sukarsa, and A. A. K. A. C. W, “Aplikasi Sistem Reminder Masa Kadaluarsa Berbasis GIS dengan Platform Android,” *Merpati*, vol. 3, no. 1, pp. 31–39, 2015.

[4] M. Syani, “Perancangan Aplikasi Pemesanan Catering Berbasis Mobile Android,” *J. Ilm. Ilmu dan Teknol. Rekayasa*, vol. 1 Nomor 2, 2018.

[5] D. Ratnasari and H. F. Hadi, “Rancang bangun aplikasi penyewaan lapangan futsal berbasis android,” vol. 16, pp. 144–157, 2018.

[6] F. Kartika, S. Dewi, T. D. Indriasari, and Y. Prayogo, “Rancang Bangun Aplikasi Pengingat Kegiatan Akademik Berbasis Mobile,” pp. 303–312, 2016.

[7] K. Tone, “Rancang Bangun Aplikasi Class Reminder Berbasis Android,” vol. 3 nomor.1, no. April, 2018.

[8] A. R. Pahlevi, N. Ismawati, and M. Cs, “Analisa Perancangan Aplikasi Musik dengan Reminder Event Berbasis Android,” vol. 1, no. 5, pp. 159–166, 2019.

[9] V. No, M. N. K. Nababan, R. S. Putra, and N. A. D. Hutagaol, “Aplikasi Pemesanan Kamar Hotel Berbasis Android,” vol. 2, no. 2, pp. 45–52, 2019.

[10] F. Panca sari, “Aplikasi Sistem Informasi Pemesanan Koki dan Masakan Rumahan Berbasis Android,” vol. vol 1 no 2, no. Desember, 2018.

[11] S. S. Nate, P. S. Navele, V. B. Mote, and P. L. S. Naik, “Smart Reminder Application With Gps System,” pp. 131–134, 2016.

[12] P. V. B. Dhore, S. Thakar, P. Kulkarni, and R. Thorat, “Digital Table Booking and Food Ordering System Using Android Application,” vol. 2, no. 7, pp. 76–81, 2014.

[13] A. F. Wali and L. T. Wright, “Customer relationship management and service quality: Influences in higher education,” no. July, 2016.

[14] S. S. L. S, “Online Token Booking Application,” vol. 38, no. 6, pp. 297–301, 2016.

[15] S. Barjtya, A. Sharma, and U. Rani, “A detailed study of Software Development Life Cycle (SDLC) Models,” vol. 6, no. 7, pp. 22097–22100, 2017.

[16] Wilieyam and G. N. Sevani, “Aplikasi reminder pengobatan pasien berbasis sms gateway,” *Inkom*, vol. 7, no. 1, pp. 13–20, 2013.

[17] A. Sonita and R. F. Fardianitama, “APLIKASI E - ORDER MENGGUNAKAN FIREBASE DAN ALGORITME KNUTH,” vol. V, no. September, pp.

38–45, 2018.

[18] E. Masiello and J. Friedmann, *Mastering react native*. Packt Publishing Ltd, 2017.

[19] A. Budiyanto, “Pengantar Cloud Computing,” *Cloud Indones. Jakarta*, pp. 1–10, 2012.

International Journal of Computer Applications Technology and Research
Volume 9–Issue 02, 53-58, 2020, ISSN:-2319–8656

Geographic Information System for Booking Beauty Salon and Barber Shop with an Android-Based E-CRM Approach

I Kadek Dharma Krisna Putra
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

I Nyoman Piarsa
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

I Made Sukarsa
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

Abstract : Beauty salons and barber shop are a necessity for almost everyone. Nowadays, the business processes of them mostly still use conventional methods. The method gives obstacles to customers who have a lot of activities, for example, they have to come directly to the beauty salon or barber shop to take the queue. In addition, it is hard to promote, communicate and assess without E-CRM media. The location is also difficult to find because there is no guide mark on Google Maps. The Geographic Information System for Booking Beauty Salon and Barber Shop provides solutions for making scheduled bookings and transactions by using the e-CRM approach as in the application of promotional features to attract customers. There are some features offered in the application, such as chat features for communication, rating and review features to assess the services obtained, and route feature as a solution to show the location of selected vendors. The implementation process requires an Android smartphone and a computer device with software including Android Studio, SQLYog, XAMPP, and Visual Studio Code. The results of the implementation is the system can be successfully applied based on experiments and testing directly.

Keywords: Geographic Information Systems, Booking, Promotion, E-CRM, Android

1. INTRODUCTION

Beauty salon is a place for hair care and makeup that generally serves female customers, while barbershop is a place for hair, moustache or beard care that serves male customers. The numbers of service providers make it difficult for customers to make choices in finding the suitable one. This is also caused by not establishing a good relationship between service providers and service users, for example, incompatibility of promotional services offered, service quality, and price offered. This need is trying to be accommodated by providing information systems to facilitate interactions between users and service providers that include bookings, promotions, transactions, chatting features that can be accessed using an Android device. This system is implemented by an E-CRM approach that prioritizes customer satisfaction and loyalty. There are several features offered by this system, for example the implementation of a payment system with an electronic wallet model that can be refilled as a means of payment, the availability of chat features to communicate directly with service providers, there are rating and review features to get an overview of service quality provided or as an assessment of services provided, and customers can also use promotions provided by service providers.

There are several studies used as references in the recent study to produce solutions related to online ordering, GIS, and CRM. A research conducted by I Putu Warma Putra produced an application that facilitates customers in ordering a taxi and assisting the driver in picking up a customer. Some of the steps conducted are system analysis, system design, and system implementation. [1]

Yuwono, Aribowo and Setyawan conducted a research that developed a geographic information system for tourism in the Magelang area based on Android. The system can be connected directly with Google Maps. The result is the system can provide information about tourism location in the Magelang area easily because it can be operated wherever the user's location is. Therefore, this result is in accordance with the purpose of the developed application. Overall, the application successfully provides tourism information and the best tourism location from the user's position. [2]

Yunefri, Devega and Kristanto developed a web-based GIS application with the aim of providing information about culinary in Pekanbaru based on the user’s location. The development of the application is carried out by using the PHP programming language and the Harversine method to determine the closest distance of a restaurant. The results of the obtained data were tested by using a black box to show that a test by using this application is similar to the manual method. [3]

A research conducted by Neene and Kabemba is the expansion of a property mapping application with a geographic information system that can be used by local authorities in developing countries. The result of this study showed this application did not require the acquisition of attribute, spatial, and real-time image data from properties. The study was strengthened by conducting a survey on the Kafue local authority to meet the requirements needed by the system. After designing and modelling, the developed system was tested on the field with the results of 10 properties were successfully mapped. [4]

A research conducted by Kusnawi discussed ordering problems that cause inconvenience to customers, such as running out of seats and also uncertainty. This problem is trying to be solved by providing a table and food ordering information system and placing it on the main server of a desktop-based restaurant. The results can be developed into web and desktop applications. The desktop application is a medium to verify online orders made by customers via the web, therefore, different platforms orders can still be entered and verified. This system supports the use of balances that can be topped up as a means of payment. [5]

Aulia Aulia, Zakir, Dafitri, Siregar, & Hasdiana conducted a research that related to the situation in the restaurant. This study designed a system that can speed up ordering and food processing in restaurants. The results of this study also allow ordering data to be sent over a wireless network that connects smartphones to computers in the kitchen. As the result, the order can be directly read by the chef because it is shown on the LCD screen. [6]

A research conducted by Rosadi and Andriawan developed an Android-based system that can facilitate in finding boarding houses from certain areas and also promote boarding houses found in the city of Bandung. The methods used in this research are descriptive analysis and development, data collection methods, and object-oriented system development methods. The results of this study indicate that an application for a boarding house can facilitate users in obtaining information, therefore the users can find boarding houses easily and according to their criteria. [7]

A research conducted by Mila Afrina and Ali Ibrahim used E-CRM concept. It resulted in a strategy to obtain, consolidate and analyze data to be used to interact with customers (students, students, teachers, lecturers, and the community). Thus, a comprehensive view of the customer and a better relationship with the customer are created. [8]

A research conducted by Le Tan discusses the progress of information technology. Thereby, it drives changes in consumer behaviour in shopping, especially in e-commerce. The results obtained are indicators of the success of e-CRM implementation in e-commerce companies. This research uses a descriptive method with the Library Research approach by looking at several previous journals and using three indicators: customer complaints, customer loyalty and management control of e-CRM implementation. The result is that these indicators can be used as a benchmark in determining the level of success in implementing e-CRM in an e-commerce. [9]

A research conducted by Tamara Luarasi, Andi Domi, Tomi Thomo, Agim Kasaj, and Ergon Baboci explained that the latest technology has supported the latest business scenarios and the application of existing ones. Its application has been tested on the livestock market, which is in dire need of the presence of technology especially in difficult rural zones. This study represents the method used in making digital brokers. [10]

Neny Rosmawarni developed a system that provides recommendations for application development. This study uses a collaborative filtering method, where the data is taken based on user feedback such as reviews and ratings. This application is an Android-based and was tested in the field of beauty like a beauty salon. With some recommendations from this application, the users can consider a beauty salon that matches their needs, budget and the suitability of the type of care provided based on the needs of the user. [11]

An information system with the implementation of web-based customer relationship management (CRM) is the final result of a research conducted by Tukino. It can be used to manage customer complaints with PT Indoritel Makmur International Tbk (Indomaret) Batam as the case study. The method used is Extreme Programming which is most widely used approach for developing this software. The results of this study are CRM information systems with a web platform [12]

Irvan Prastya, Sarip Hidayatuloh and Nidaul Hasanati conducted a research by designing a website-based E-CRM information system at PT Persada Duta Beliton. It can be used for maintaining company relationships with customers. The results of this study are the system has a tour package booking feature, a chat feature for customer communication and a feature for giving tours package reviews based on customer experience. In addition, there is a promotional media for the customer, such as congratulating a celebration by giving gifts in the form of discounts to them. The model in developing the system is the Rapid Application System (RAD). The system design is conducted with UML. [13]

Anharudin, Donny Fernando and Novi Khristina Putri developed an E-Booking Information System that can be used to facilitate customers to order karaoke rooms at Happy Puppy in Cilegon City. The application is developed and designed by using UML modelling. In addition, it is created with the PHP programming language and uses MySQL to store the database. The existence of an E-Booking Information System is expected to help customers in ordering karaoke rooms at Happy Puppy in Cilegon City, Banten. [14]

A research conducted by Devi Mawarni and Rinabi Tanamal developed an information system for a medium scale beauty salon by using Visual Basic 2010 programming with SDLC waterfall. The study was conducted by interviews to obtain and conduct data analysis. The implementation resulted in a program that could solve the problems found in medium-scale beauty salons. [15]

Research conducted by Luh Gede Sri Handayani, I Nyoman Piarsa, Kadek Suar Wibawa produced a geographic information system for web-based village road mapping. This system using Google Maps, with the polyline feature making it possible to describe the road network and geometry library that can calculate road lengths. Data collection on this system is done by two ways, digitizing and input coordinates, and processing data on master data. The results of this mapping provide information of road names, types of road surfaces, road lengths and road conditions. [16]

Research conducted by I Made Widnyana, I Nyoman Piarsa, A. A. K. Agung Cahyawan W developed a geographic information system to map the location of workshops in Denpasar. This system can be accessed using an Android smartphone with Google Maps API support that allows user to find out information about the workshop. The results of this study are the application can allow users to get route information with the direction feature. User can see detailed information about services and spare parts provided by the workshop. For workshop owners can update service information and create promos. [17]

Research by I Wayan Wahyu Gautama, I Ketut Gede Darma Putra, I Made Sukarsa produced a geographic information system for mapping coastal tourism objects in southern Bali. This application is designed using the Google Maps API which can be run on android devices. The test results obtained a percentage of 80% with a very good value in terms of the presentation of content and beach information, this application can also display the route to the beach. [18]

Based on the background and references above, this research focuses on developing a system that facilitates interaction between the service providers of beauty salon and barber shop with customers. There is a difference of the current study with the previous one. In the previous study, there has been no solution that facilitates salon and barber users. It is because the previous study only focused on one vendor, unscheduled bookings, and the unavailability of e-CRM features such as promotions, chat, ratings and reviews. There are some topics that still be used as supporting references in building a beauty salon and barbershop information system such as E-CRM, booking or booking, using Google maps in GIS and so on. This research is also expected to develop a friendly system for various vendors.

2. LITERATURE REVIEW

The literature Review discusses about supporting theories in conducting the study, such as Android, Android Studio, Geographic Information System (GIS), E-CRM and Database.

2.1 Android

Android is a software that is commonly used on smartphone devices which includes operating systems, middleware and key applications. Android-based application development can be performed by using the Java and Kotlin programming languages. Android OS has core applications including SMS programs, contacts, browsers, maps, calendars, e-mails and others. Android is an open source platform, therefore developers are free to participate in developing applications for Android devices that are great in capabilities and innovative. The android developers are also given the freedom to access hardware information, access location information, run background services applications, set alarms, make notifications to the status bar, and many more.

2.2 Android Studio

Android Studio is an Integrated Development Environment (IDE) that is used to develop applications based on the Android operating system and created with the IDB's JetBrains IntelliJ software. IDE makes it easier for android application developers. In addition, it replaces Eclipse Android Development Tools (ADT) which were previously the main IDE for Android application development. Android studio was announced for the first time at the Google I / O conference on May 16, 2013. On May 2013, it was still a preview of version 0.1 until it was released in June 2014 where it had entered the beta stage with version 0.8. The stable version was finally released in December 2014, with version number 1.0.

2.3 Geographic Information System (GIS)

Geographical Information System (GIS) is an information system that can store, process, analyze, and provide information based on user's geographic references. It is a system for utilizing and producing spatial data process and analysis as well as non-spatial data in obtaining various information related to spatial aspects, both scientific, commercial, management and policy oriented.

2.4 E-CRM

E-CRM is a business strategy that uses information technology that gives companies a broad, reliable and integrated customer view. In other words, all customer processes and interactions can be used to be processed and analyzed with the aim of finding strategies to maintain and expand profitable relationships simultaneously. Furthermore, E-CRM is a powerful and flexible platform intended to facilitate customers in interacting with companies. It is also a CRM that is applied electronically by using media such as web browsers, the internet and electronic media such as e-mail, call centers. Sometimes, it is also referred as E-service which means electronic services for customers without meeting in person. The concept of E-CRM is ‘sense’ and ‘respond marketing’ which means good relations with consumers will be realized if a company understands customer desires and uses them as opportunities as well as responds positively to consumer desires.

E-CRM is currently needed in order to save costs and improve the efficiency of managing customer relationships. It is believed that it has been used since the mid-1990s. At that time, the customers have started it by using web browsers, the internet and electronic transaction media, such as e-mail. E-CRM is a CRM that is enhanced by using internet media.

2.5 Database

A database is a collection of data stored in a computer. The data can be processed, modified or manipulated with the help of software that aims to produce information. The definition of the database includes specifications, such as data types and data structures. In addition, there are also restrictions on the data to be stored. The database is an important aspect in information systems because it functions as a data storage repository. Furthermore, it is important because it has a role in organizing data, avoiding duplication, preventing the existence of unclear relationships between data. In processing the data to and from data storage media, it requires the help of software called DBMS or Database Management System. It is a software that makes database users easier to process, control and access the data practically and efficiently.

3. RESEARCH METHODS

Beauty Salon and Barber Shop Information Systems use waterfall-based lifecycle development software in designing systems. The SDLC model with the waterfall method has 5 stages in developing software, including analysis, design, implementation, testing and maintenance. These stages can be illustrated through Figure 1.

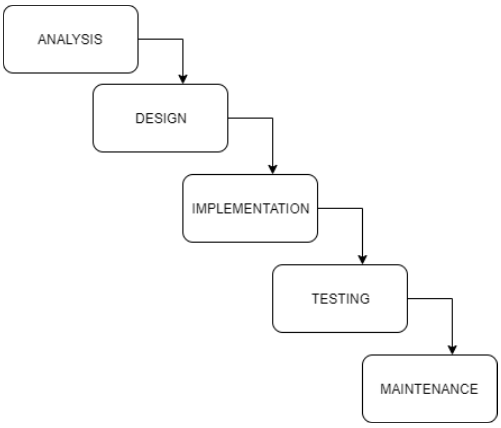


Figure 1. Stages of SDLC Waterfall

Figure 1 is the stage of system development by using the SDLC (System Development Life Cycle) waterfall method. It consists of 5 stages. Analysis is the first stage that is performed to be able to determine the needs in building and designing systems. This stage will answer all the needs of both the vendor and the customer. The second stage is designing the user interface of the beauty salon and barber shop information system. The third stage is implementing the information systems according to the needs of customers and vendors. The implementation requires several hardware devices such as smartphones and computers, while the software requires Android OS, Visual Studio Code, Android Studio, XAMPP and Windows 10. The fourth step is testing the system to find out the performance of it. The test was conducted to minimize errors or bugs contained in the system. If that problem happened, it can be fixed or redesigned. The fifth stage is the process of maintaining the system performance both from the internal system and errors caused by use.

3.1 General Overview of the System

The general overview of the Salon and Barber Shop Geographic Information System with the Android-based E-CRM approach is as shown on figure 2.

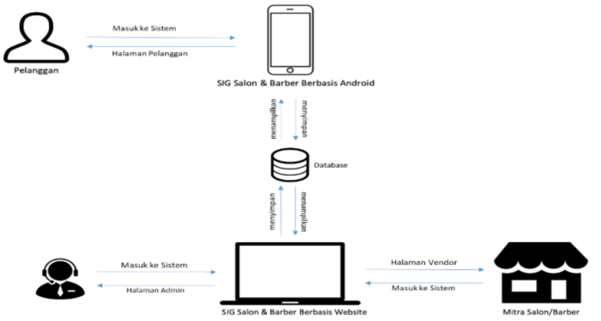


Figure 2. System’s General Overview

Figure 2 is a general overview of the system where there are three actors whose interaction is facilitated by the system. Based on the figure, the customer can enter the system on an Android smartphone device to book a salon or barber service. The system stores information from the customers to the database. Furthermore, it informs the admin and the vendor based on the orders placed and in accordance with the transaction method chosen by the customer. Every customer’s order will have a unique ID in order to identify the actor who has a role in confirming the order and storing detailed information of the order placed such as date and time, price, order list and others. The vendors use the system to manage vendor data, add vendor services, confirm top up balances and transactions through them. The admin is the actor who has the most extensive control of the system. They have the authority, such as accepting registration and confirmation of new vendors, managing vendor and customer data, managing transactions by using the payment transfer and balance method, managing top up with the transfer method, and managing promos.

3.2 Context Diagram

Diagram context is a diagram that consists of a process that represents the whole system. It illustrates the input or output of a system. The context diagram of the beauty salon and barber shop information system is explained with Figure 3.

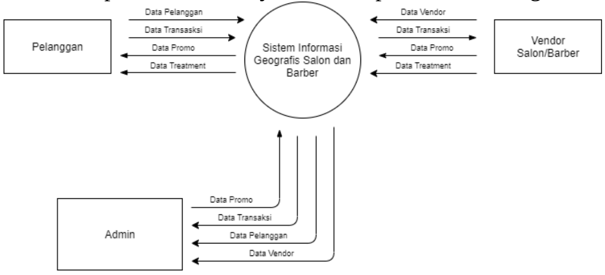


Figure 3. Context Diagram of Sabrina

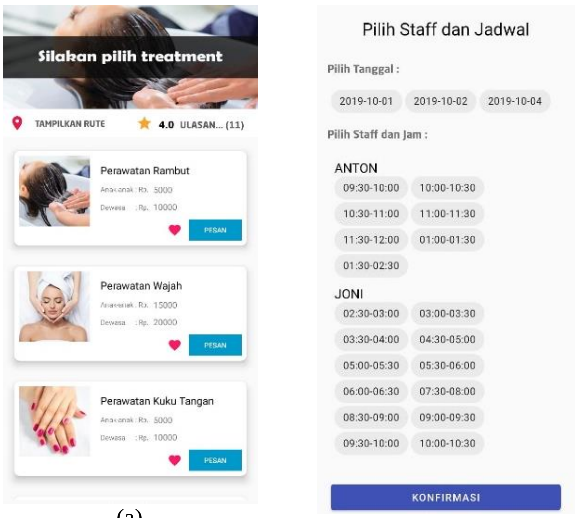
Based on Figure 3, it can be seen that the beauty salon and barber shop information system process involves three external entities namely admin, customer, and vendor. The vendor is a service provider of the system. The system will provide and request information such as vendor data, promo data, transaction data, and treatment data provided by the beauty salon or barber shop. Customers can fill out personal data, make transactions, receive promo data and treatment contained in the vendor. Admin manages the process of Beauty Salon and Barber Shop Information system. In addition, it plays a role as a bridge of the transaction process between the beauty salon / barber shop vendor and the user.

4. RESULT AND DISCUSSION

The results and discussion on the application of the beauty salon and barber shop information systems will be explained in the form of booking feature, promotion and E-CRM feature testing.

4.1 Booking Feature

Customers can choose the treatment from each of barbershop and choose the desired time to get the service they ordered. The testing Booking feature can be seen in Figure 4.



Perawatan Rambut
Anak-anak : Rp. 5000
Dewasa : Rp. 10000

Perawatan Wajah
Anak-anak : Rp. 15000
Dewasa : Rp. 20000

Perawatan Kuku Tangan
Anak-anak : Rp. 5000
Dewasa : Rp. 10000

Pilih Staff dan Jadwal

Pilih Tanggal :
2019-10-01 2019-10-02 2019-10-04

Pilih Staff dan Jam :

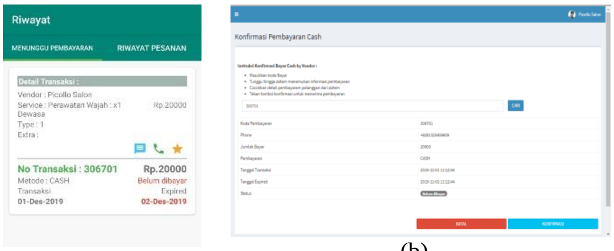
ANTON
09:30-10:00 10:00-10:30
10:30-11:00 11:00-11:30
11:30-12:00 01:00-01:30
01:30-02:30

JONI
02:30-03:00 03:00-03:30
03:30-04:00 04:30-05:00
05:00-05:30 05:30-06:00
06:00-06:30 07:30-08:00
08:30-09:00 09:00-09:30
09:30-10:00 10:00-10:30

(a) (b)

Figure 4. List of salons and treatment selection in booking process.

The customer can choose the treatment from each of registered beauty salon or barber shop in the system as shown in figure 4. (a). Then, the customer can determine the time when they want to get the treatment and employee information from the serving vendor as shown in figure 4. (b). The confirmation stage for incoming orders can be seen in Figure 5.



Detail Transaksi :
Vendor : Picollo Salon
Service : Perawatan Wajah : x1 Dewasa Rp.20000
Type : 1
Extra :

No Transaksi : 306701 Rp.20000
Metode : CASH Belum dibayar
Transaksi 01-Des-2019 Expired 02-Des-2019

Konfirmasi Pembayaran Cash

Instruksi Konfirmasi Bayar Cash by Vendor :
- Masukkan kode Bayar
- Tunggu hingga sistem menemukan informasi pembayaran
- Cocokkan detail pembayaran pelanggan dari sistem
- Tekan tombol konfirmasi untuk menerima pembayaran

306701

Kode Pembayaran	306701
Phone	[illegible]
Jumlah Bayar	20000
Pembayaran	CASH
Tanggal Transaksi	2019-12-01 11:12:44
Tanggal Expired	2019-12-02 11:12:44
Status	Belum dibayar

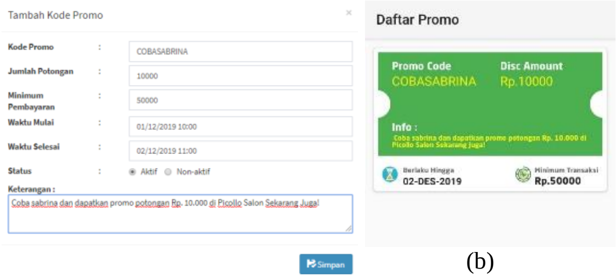
(a) (b)

Figure 5. Customer’s booking list and vendor’s payment confirmation

Figure 5. (a) is a display of history when a customer has placed an order. The history shows a list of treatments that have not been paid yet and confirmed by the vendor. Figure 5. (b) is a list of orders that will be confirmed by the vendor, therefore the customer can enjoy the treatment.

4.2 Promotion Feature and E-CRM

The promotion feature is part of implementing E-CRM on the system. Beside promotion feature, there are chat feature as well as ratings and reviews feature which are also part of the implementation of E-CRM. The testing of these features can be seen in Figure 6 and Figure 7 below.



Tambah Kode Promo

Kode Promo : COBASABRINA
Jumlah Potongan : 10000
Minimum Pembayaran : 50000
Waktu Mulai : 01/12/2019 10:00
Waktu Selesai : 02/12/2019 11:00
Status : ● Aktif ○ Non-aktif
Keterangan : Coba sabrina dan dapatkan promo potongan Rp. 10.000 di Picollo Salon Sekarang Juga!

Simpan

Daftar Promo

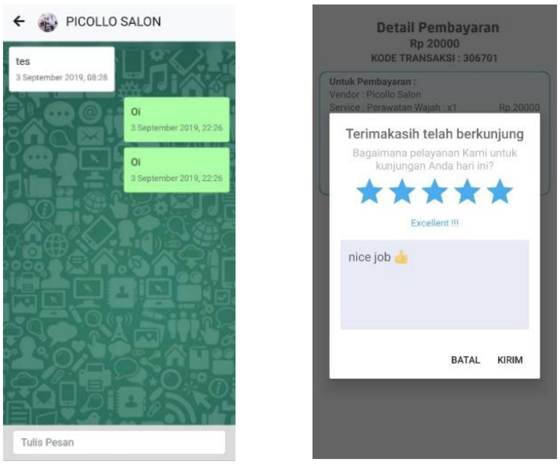
Promo Code COBASABRINA
Disc Amount Rp.10000
Info : Coba sabrina dan dapatkan promo potongan Rp. 10.000 di Picollo Salon Sekarang Juga!
Berlaku Hingga 02-DES-2019
Minimum Transaksi Rp.50000

(b)

(a)

Figure 6. Vendor create a promo and customer checking the promo

Figure (a) is a feature provided for vendors in creating promotional codes. The promotional codes can be used by customers on the promo page as shown in figure (b).



PICOLLO SALON

tes
3 September 2019, 08:28

Oi
3 September 2019, 22:26

Oi
3 September 2019, 22:26

Tulis Pesan

Detail Pembayaran
Rp 20000
KODE TRANSAKSI : 306701

Untuk Pembayaran :
Vendor : Picollo Salon
Service : Perawatan Wajah : x1 Rp.20000

Terimakasih telah berkunjung
Bagaimana pelayanan Kami untuk kunjungan Anda hari ini?
★★★★★
Excellent !!!

nice job 👍

BATAL KIRIM

(a) (b)

Figure 7. Chatting feature and Giving a review and rating

Figure 7. (a) is a chatting page provided as a communication tool between customers and vendors. Figure 7. (b) is ratings and reviews from the customers of services provided by vendors.

5. CONCLUSION

The Android Based High School Management Information The conclusions that can be drawn from implementing the system are as follows:

First, the design of beauty salon and barber shop information systems is conducted by making a general picture design, data flow diagrams, ERD and PDM as the basic structure in developing a database system. Second, the stages of the implementation of the information systems are using the Java programming language assisted by the Android Studio IDE. They were used in order to develop customer-side applications on the Android platform and PHP programming language to develop a web platform for the admin and vendor in dealing with customers. The testing process is conducted by testing both the Android platform application and the website platform application directly. Third, the results of the questionnaire show the measurement of E-CRM performance towards the Android-Based Beauty Salon and Barber Shop Geographic Information System which can be stated to be running well from the vendor and customer side.

REFERENCES

[1] I. W. G. N. I Putu Warma Putra, “Rancang Bangun Sistem Informasi Geografis Pemesanan Taksi Berbasis Android,” *J. Sist. dan Inform.*, pp. 50–59, 2014.

[2] B. Yuwono, A. S. Aribowo, and F. A. Setyawan, “Sistem Informasi Geografis Berbasis Android Untuk Pariwisata Di Daerah Magelang,” *J. Ilm. Tek. Inf.*, vol. 2015, no. 2015, pp. 68–74, 2015.

[3] Y. Yunefri, M. Devega, and D. Kristanto, “Geographic Information System (Gis) for Culinary in Pekanbaru using Herversine Formula Geographic Information System (Gis) for Culinary in Pekanbaru using Herversine Formula,” 2017.

[4] V. Neene and M. Kabemba, “Development of a Mobile GIS Property Mapping Application using Mobile Cloud Computing,” *IJACSA) Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 10, pp. 57–66, 2017.

[5] Kusnawi, “Perancangan Sistem Informasi Pemesanan Meja Dan Makanan (Studi Kasus Restoran Abc),” *J. Dasi*, vol. 14, no. 1, pp. 40–43, 2013.

[6] R. Aulia, A. Zakir, H. Dafitri, D. Siregar, and H. Hasdiana, “Mechanism of Food Ordering in A Restaurant Using Android Technology,” *J. Phys. Conf. Ser.*, vol. 930, no. 1, 2017.

[7] D. Rosadi and F. O. Andriawan, “Aplikasi sistem informasi pencarian tempat kos di kota bandung berbasis android,” *J. Comput. Bisnis*, vol. 10, no. 1, pp. 50–58, 2016.

[8] M. Afrina *et al.*, “Rancang Bangun Electronic Costumer Relationship Management (E-CRM) Sebagai Sistem Informasi Dalam Peningkatan Layanan Perpustakaan Digital Fakultas Ilmu Komputer Unsri.,” *J. Sist. Inf.*, vol. 5, no. 2, pp. 629–644, 2013.

[9] T. Le Tan, “Successful Factors of Implementation Electronic Customer Relationship Management (e-CRM) on E-commerce Company,” *Am. J. Softw. Eng. Appl.*, vol. 6, no. 5, p. 121, 2017.

[10] T. Luarasi, A. Domi, T. Thomo, A. Kasaj, and E. Baboci, “Cloud based communication in B2B model,” *Proc. - Asia Model. Symp. 2014 8th Asia Int. Conf. Math. Model. Comput. Simulation, AMS 2014*, pp. 27–32, 2014.

[11] N. Rosmawarni, “Perancangan sistem rekomendasi untuk pengembangan aplikasi salon terpadu berbasis android,” *Rekayasa Inf.*, vol. 6, no. 1, pp. 61–70, 2017.

[12] Tukino, “Rancang Bangun Sistem Informasi Customer Relationship Management (Crm) Berbasis Web,” *Comput. Based Inf. Syst. J.*, vol. 06, no. 01, pp. 12–22, 2018.

[13] I. Prastya, S. Hidayatuloh, and N. Hasanati, “RANCANG BANGUN i-CRM (INTERACTIVE CUSTOMER RELATIONSHIP MANAGEMENT) UNTUK JASA AGEN PERJALANAN WISATA (Studi Kasus: PT Persada Duta Beliton),” *Stud. Inform. J. Sist. Inf.*, vol. 10, no. 1, pp. 45–52, 2018.

[14] Fandhilah, Dany Pratmanto, A. Fatakhudin, “Rancang Bangun Sistem Informasi E-Booking Ruang Karaoke Berbasis Web (Studi Kasus : Karaoke Keluarga Happy Puppy),” *Indones. J. Softw. Eng.*, vol. 3, no. 2, pp. 68–76, 2018.

[15] Devi Mawarni; Rinabi Tanamal, "Rancang Bangun Sistem Informasi pada Salon Skala Menengah," *Inform. dan Sist. Inf. ISSN 2460- 1306*, vol. 02, no. 01, pp. 74–81, 2016.

[16] L. G. Sri Handayani, I. N. Piarsa, and K. Suar Wibawa, "Sistem Informasi Geografis Pemetaan Jalan Desa Berbasis Web," *CSRID (Computer Sci. Res. Its Dev. Journal)*, vol. 6, no. 3, p. 171, 2015.

[17] I. Widnyana, I. Piarsa, and A. Agung Cahyawan W., "Aplikasi Sistem Informasi Geografis Bengkel di Kota Denpasar Berbasis Android," *Merpati*, vol. 3, no. 1, pp. 23–30, 2016.

[18] I. W. W. Gautama, I. K. G. D. Putra, and I. M. Sukarsa, "Aplikasi Pemetaan Objek Wisata Pantai Bali Selatan Berbasis Android," *Merpati*, vol. 4, no. 1, pp. 43–51, 2016.

International Journal of Computer Applications Technology and Research
Volume 9–Issue 02, 59-64, 2020, ISSN:-2319–8656

Network Security Monitoring System on Snort with Bot Telegram as a Notification

I Made Ari Sulistya
Departement of Information Technology
Faculty of Engineering Udayana University
Bukit Jimbaran, Bali, Indonesia

Gusti Made Arya Sasmita
Departement of Information Technologiy
Faculty of Engineering Udayana University
Bukit Jimbaran, Bali, Indonesia

Abstract: Network security in the digital era needs more attention. IDS (Intrusion Detection System) is one of the anticipation method that can be used to protect computer server. Snort IDS only comes with terminal notification or web based, this method has a weakness which to transfer information to administrator network directly. Telegram is an open source instant messaging application. Combination between those applications, produce a perfect transformation to administrator network directly through smartphone. Contrive and testing are the best methods to build this network monitoring system. The Prevention methods is added to support this network monitoring system. Penetration testing is divided by two different types such as, DDoS and port scanning. The result of those two types penetration testing show that Snort IDS is succeeded to detect those tests. The time different between Snort detection and Bot Telegram after ten times attempt in sending messages is 77,1 seconds for Snort detection and 4,05 seconds for Bot Telegram. The time different between two types of penetration after ten times attempt is 6,1 seconds for DDoS UDP and 2 seconds for Nmap portscan.
.

Keywords: IDS (Intrusion Detection System), Snort, Telegram, Penetration Testing

1. INTRODUCTION

In the last five years, cybercrime is increasing which includes identity theft, viruses, and system intrusions. Therefore, Intrusion Detection System (IDS) took an important role in detecting intrusions so that can be addressed immediately. Snort is one of the leading Network-Based IDS (Intrusion Detection System) software with nearly four hundred thousand users [2]. However, Snort does not provide a sufficient GUI (Graphical User Interface) so that the user has to install another application separately such as BASE to get a better GUI [3].

IDS alert system with web based interface like BASE cannot offer notify to the system administrator, it is possible that users may miss some attacs so that the response become too late to do [1]. The IDS (Intrusion Detection System) with real time notification system is highly needed [6].

The growing of Instant Messenger in this era, not only used in desktop. Instant Messenger has been commonly used in mobile devices. Instant Messenger also used in many platforms as real time notification system. Instant Messenger provides query facility that cannot be easily applied in other communication media [3]. System administrator are able to interact with the system to get the status and condition of the system even able to modify.

Instant Messenger only needs an internet connection to connect to the server. Meanwhile, SMS gateway requires cost of each sent message so applying SMS gateway in a system notification need no small cost. So the costs incurred when using Instant Messenger is much less [4].

Telegram Messenger is widely used instant messaging protocol with about 200 milion users that runs on many mobile operating systems such Android, iOS, and Windows phone. Telegram Messenger also can be found on desktop operation system such as Windows, Linux, aand MacOS [5]. Telegram Messenger also a platform messenger that easily to use with many modification to synchronize with any application. Based on the facts above, we utilize Telegram Messenger as an interactive interface for Snort IDS with real time notification [2]. The function of this messenger application can obtain intrusion alerts and their detail information in real time manner, also can do small prevention system to protect the system.

2. METHODOLOGY RESEARCH

The research was done based on Snort structure and integration between Telegram Messenger. The design of this system will monitor all types of suspicious data packages and the flow of data that enters the computer server through Snort. Here is a scheme of Intrusion Detection System Snort with BASE as web interface [2].

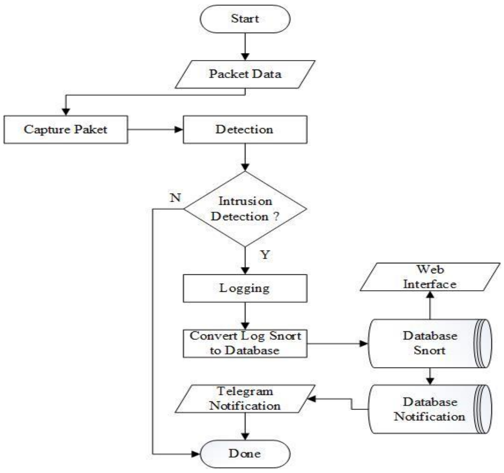


Figure. 1 Flowchart of Intrusion Detection System Snort

Figure 1 is flowchart design of Intrusion Detection System Snort with BASE as web interface. IDS console read each network packet from the target [5]. The packet is then inspected wheter the packets is a malicious packet. Each detection result will be stored into the log. The third party application called barnyard2 will convert Snort log into database and stored to system administrator using web interface or database notification of Telegram Messenger [4].

2.1 Telegram Messenger API

Telegram Messenger provides an API (Application Program Interface) that allow developers to build application integrated to Telegram Messenger. Telegram Messenger uses bot to communicate with the system server [1]. User must registration their own bot to perform authorization to perform activities with Telegram Messenger such send messages or retrive messages. Here is a cheme flowchart design to create our own bot Telegram.

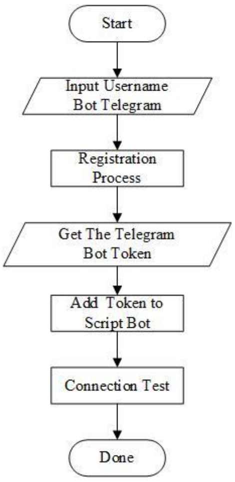


Figure. 2 Flowchart design Registration Bot Telegram

User must intall Telegram on the devices and type code to create bot on telegram. Input the username of bot application, the Telegram server will reply with bot token to connect the script of our bot to the server Telegram.

Integration between Telegram Messenger with Snort scheme can be use after this registration bot. Create database notification to accommodate Snort malicious packet information. Here is the design scheme of integrated Snort with Telegram Messenger Notification.

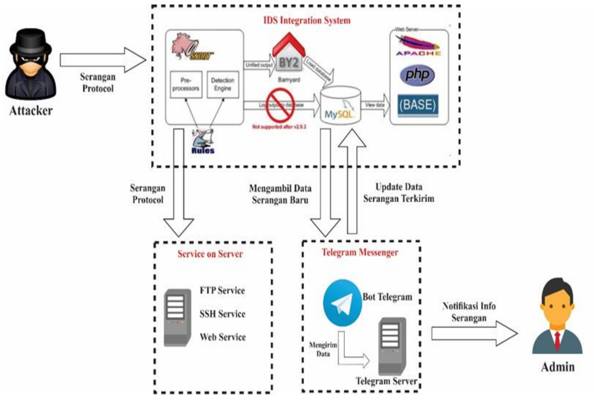


Figure. 3 Scheme design Snort with Telegram Notification

The design of the Snort with Telegram Notification is made from the detection of packages suspected by Snort and then entered into the database by a third party application called barnyard2. Telegram Messenger bot works according to the scripts to send notification to system administrator.

3. CONCEPTS AND THEORIES

Literature review contains supporting theories in the research that will be conducted. The theories including Intrusion Detection System and Telegram Messenger bot will be discussed as follows.

3.1 Intrusion Detection System

Intrusion Detection System (IDS) is a system that monitors network traffic for suspicious activity and issues alerts when such activity is discovered. IDS is a software application that scans a network or a system for harmful activity or policy breaching [6].

IDS types range in scope from single computers to large networks. The most common classifications are network intrusion detection systems (NIDS) and host-based intrusion detection systems (HIDS). System that monitors important operating system files is an example of an HIDS, while a system that analyzes incoming network traffic is an example of an NIDS [3]. IDS classification it is possible too by detection approach. The most well-known variants are signature-based detection (recognizing bad patterns, such as malware) and anomaly-based detection (detecting deviations from a model of good traffic, which often relies on machine learning). Another common variant is reputation-based detection (recognizing the potential threat according to the reputation scores). Some IDS products have the ability to respond to detected intrusions. Systems with response capabilities are typically referred to as an intrusion prevention system. Intrusion detection systems can also serve specific purposes by augmenting them with custom tools, such as using a honeypot to attract and characterize malicious traffic [1].

Snort employs both signature based techniques and anomaly based techniques to detect an intrusion. Signatures are used for detecting intrusions. Snort has a rich rule set which depend upon the signatures present in either the header part of the packet or payload of the packet so as to detect intrusions [2].

3.2 Snort

Snort is a light-weight intrusion detection tool which logs the packets coming through the network and analyzes the packets. Snort checks the packets coming against the rules written by the user and generate alerts if there are any matches found. The rules are written by the user in a text file which is linked with snort.conf file where all the snort configurations are mentioned. There are few commands which is used to get snort running so that it can analyze network behavior [4]. The architecture of snort can be categorized into five basic modules namely Libcap, Packet Decoder, Preprocessors, Detection Engine and Output plugins.

The traffic comes from Internet are received by routers and passed to switch. The switch then delivers this data traffic to firewalls for first level of evaluation. After that the firewalls passed them to the Ethernet adapter of server. Here the Snort came into focus for any type of evaluation of those data packets [6].

There are different types of preprocessors included in snort 2.9.7.2 with optional options, however in this paper no preprocessor configured but for actual intrusion detection and prevention they are necessary to control [2].

3.2.1 Detection Engine

This portion of snort is principally very dynamic and unified. This module is very vital in terms of multiple rules examination in terms of their priority order. When snort use to inspect the packets multiple rules with different priorities are reported and stored in a queue. However then it only reports out the rules with highest priority. This is specially used to avoid deep evasion techniques if used by attacker. This makes the snort as a highly proficient in terms of attack identification [3].

3.2.2 Output Modules

This module design came up after Snort 1.6 version. This is the last segment of snort where packets come from detection engine and disseminated to network in different modes as per the convenience of the network administrator. The convenience of network administrator is in terms to view the real time alerts, logs and other parameters to evaluate the performance of the network of the organization. Third party tools such as mysql, a database, can also used for the same purpose. But in this paper the logs are stored into /var/log/snort directory [3].

4. RESULTS AND DISCUSSION

To determine the effectiveness of this system, we calculate the amount of delay between the occurrence of the incident until when the notification is received by the user. Testing was conducted using 10 similar attacks from single source of attack and using single client as the recipient of the notification.

4.1 Accuracy Rate of Time

The level of accuracy rate of time is calculated from the time of attack and detected. Average IDS detection rate depend on it. The rate of notification sent is also obtained from the difference in time sent with the time received by the admin, as shown in Table 1.

Table 1. Accuracy Rate of Time

No	Types	**Accuracy Rate of Time (Timestamp)**			
		Intruder Time (m^a,s^a)	**Snort Detection Time (m^b,s^b)**	**Telegram sent Time (m^c,s^c)**	**Admin Receive Time (m^d,s^d)**
1.	DDoS UDP	15:09:38	15:10:42	15:11:52	15:11:57
2.	DDoS UDP	15:09:49	15:10:52	15:11:54	15:11:57
3.	DDoS UDP	15:09:58	15:11:02	15:11:54	15:12:02
4.	DDoS UDP	15:10:09	15:11:12	15:11:55	15:12:02
5.	DDoS UDP	15:10:20	15:11:22	15:11:55	15:12:02
6.	DDoS UDP	15:10:28	15:11:32	15:11:55	15:12:02
7.	DDoS UDP	15:10:41	15:11:42	15:11:56	15:12:12
8.	DDoS UDP	14:48:01	14:49:04	14:49:28	14:49:30
9.	DDoS UDP	14:48:12	14:50:24	14:50:34	14:50:37
10.	DDoS UDP	14:48:22	14:50:34	14:50:39	14:50:42
11.	Nmap Portscan	11:50:51	11:53:16	11:53:19	11:53:22
12.	Nmap Portscan	11:52:28	11:55:37	11:55:41	11:55:43
13.	Nmap Portscan	11:54:31	11:56:42	11:56:45	11:56:47
14.	Nmap Portscan	11:56:11	11:57:03	11:57:06	11:57:08
15.	Nmap Portscan	11:58:02	11:58:48	11:58:50	11:58:52
16.	Nmap Portscan	12:01:22	12:02:03	12:02:05	12:02:08
17.	Nmap Portscan	12:03:05	12:03:45	12:02:48	12:02:49
18.	Nmap Portscan	12:05:35	12:06:15	12:06:18	12:06:20
19.	Nmap Portscan	12:07:46	12:08:26	12:08:29	12:08:30
20.	Nmap Portscan	12:09:17	12:10:07	12:10:10	12:10:12

Based on the Table 1. Accuracy Rate of Time measurement result above is the timestamp of two types penetration between DDoS UDP and Nmap portscan. The timestamp include of Intruder Time (m^a,s^a), Snort Detection Time (m^b,s^b), Telegram Sent Time (m^c,s^c), and Admin Receive Time (m^d,s^d). Intruder Time is timestamp that intruder start to attack the system. Snort Detection Time is timestamp that Snort detected malicious packet data. Telegram Sent Time is timestamp that bot Telegram send messages to Network administrator. Admin Receive Time is timestamp that network administrator receive all the messages of malicious packet data information.

4.2 Time Difference

Time difference between two types of penetration DDoS UDP and Nmap portscan shown in Table 2. The time difference recorded in seconds time unit. Time Difference between attacking and detection are obtained from Intruder Time (m^a,s^a), and Snort Detection Time (m^b,s^b) timestamp. Time difference between send and receive are obtained from Telegram Sent Time (m^c,s^c), and Admin Receive Time (m^d,s^d) timestamp. Here is the result of time difference shown in Table 2.

Table 2. Time Difference

No.	Types	Time Difference (Seconds)	
		Time Difference between attacking and detection (S^x)	**Time Difference Between send and receive (S^y)**
1.	DDoS UDP	64	5
2.	DDoS UDP	63	3
3.	DDoS UDP	64	8
4.	DDoS UDP	63	7
5.	DDoS UDP	62	7
6.	DDoS UDP	64	7
7.	DDoS UDP	61	16
8.	DDoS UDP	63	2
9.	DDoS UDP	132	3
10.	DDoS UDP	132	3
11.	Nmap Portscan	145	3
12.	Nmap Portscan	189	2
13.	Nmap Portscan	131	2
14.	Nmap Portscan	52	2
15.	Nmap Portscan	46	2
16.	Nmap Portscan	41	3
17.	Nmap Portscan	40	1
18.	Nmap Portscan	40	2
19.	Nmap Portscan	40	1
20.	Nmap Portscan	50	2
Amount of Time		**1542**	**81**
Average Time		**77,1**	**4,05**

Based on the table of Time Difference measurement result above the highest value between attacking and detection is 189 seconds on Nmap portscan type, while lowest value is 40 seconds on Nmap portscan type. The highest value between send and receive is 16 seconds on DDoS UDP type, while lowest value is 1 seconds on Nmap portscan type. Average Time of 20 penetration test with two types of attacking way between intruder and Snort detection is 77,1 seconds, while average time between send and receive is 4,05 seconds.

Time difference is obtained from timestamp result of Table 1. The formula to get the time difference is substraction between timestamp result on Table 1. Here is the formula to get time difference between attacking and detection.

$$S^x = ((m^b \times 60) + s^b) - ((m^a \times 60) + s^a)$$

$$S^x = (60m^b + s^b) - (60m^a + s^a)$$

Figure 4. Formula of Time Difference Attacking and Detection

Explanation :

S^x = Time Difference in seconds between attacking and detection
m^a = Minutes unit of time from attacking.
s^a = Seconds unit of time from attacking.
m^b = Minutes unit of time from Snort Detection.
s^b = Seconds unit of time from Snort Detection.

$$S^y = ((m^d \times 60) + s^d) - ((m^c \times 60) + s^c)$$

$$S^y = (60m^d + s^d) - (60m^c + s^c)$$

Figure 5. Formula of Time Difference Send and Receive

Explanation :

S^y = Time Difference in seconds between Send and Receive.
m^c = Minutes unit of time from Send.
s^c = Seconds unit of time from Send.
m^d = Minutes unit of time from Receive.
s^d = Seconds unit of time from Receive.

Analytical data set to see the graph of time difference data between sent and receive. Figure 6 shown the flow of Telegram messenger transformation rate to send messages information.

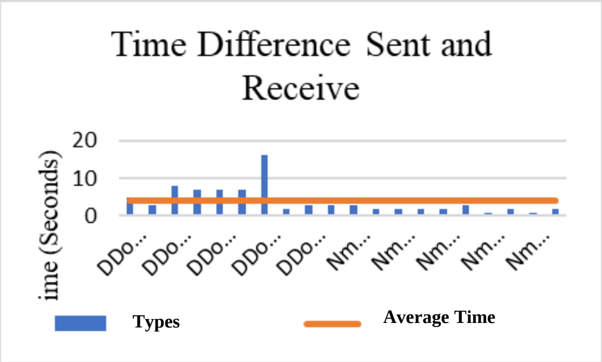


Figure 6. Analytics Chart Time Difference Sent and Receive

Figure 6 is analytical chat of time difference from send and receive messages bot Telegram. The chart shown some big difference time messages in the middle of simulation.

Analytical data set to see the graph of time difference data between Snort detection and attacking time. Figure 7 shown the flow of Snort detection rate to detected malicious packet data.

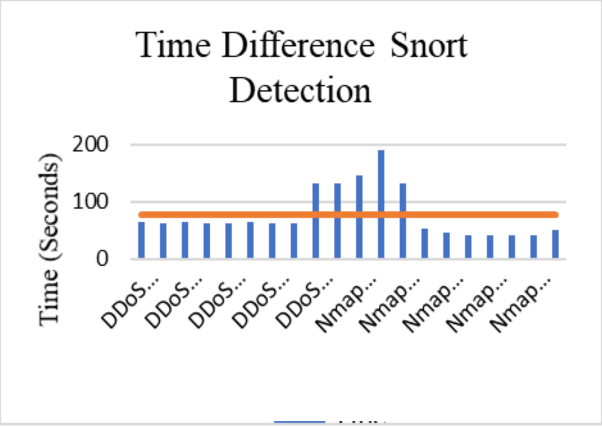


Figure 7. Analytics Chart Time Difference Snort Detection and Attacking

Figure 7 is analytical chat of time difference from Attacking and Snort Detection from detected malicious packet data. The chart shown some big difference time messages in the middle of simulation with transition betweet DDoS UDP attack to Nmap portscan attack.

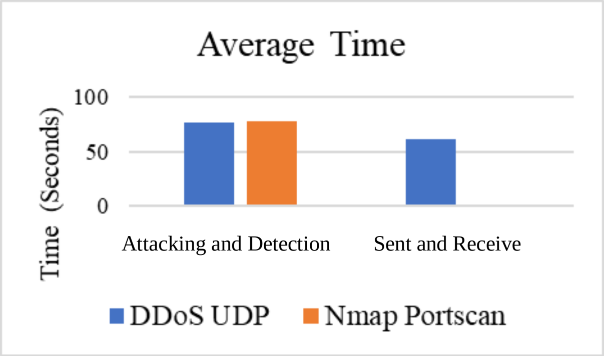


Figure 8. Analytics Chart From Average Time

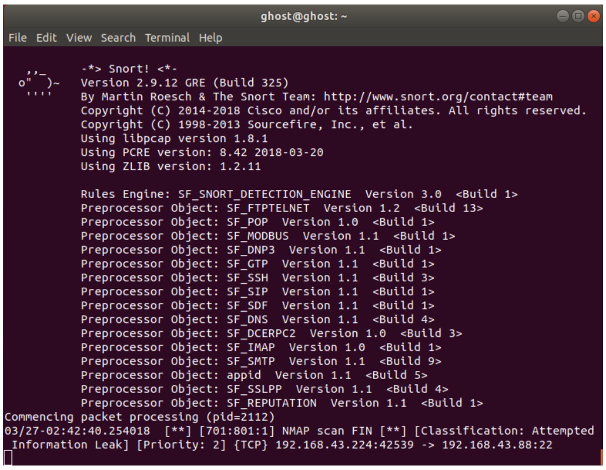
Figure 8 is analytical chat from average time of attacking and detection also sent and receive with two types of penetration in seconds unit of time. The chart of attacking and detection between two penetration test shown no big difference. The chart of sent and receive between DDoS UDP and Nmap portscan penetration shown big difference that DDoS UDP have more time to send information.

4.3 System Testing

Testing system monitoring with snort has two main part from the omputer testing and the mobile testing. This feature will be explained as follows.

4.3.1 Snort Testing

Snort that already installed on computer system must be tested wether it works or not. Snort testing can be done by run the command on terminal and setting the rules of snort to detect malicious packet data like shown on Figure 9.



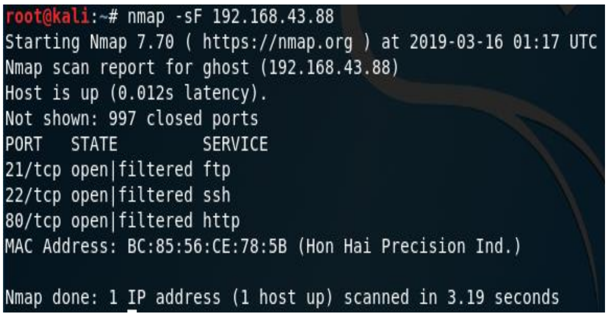
```
   ,,_     -*> Snort! <*-
  o"  )~   Version 2.9.12 GRE (Build 325)
   ''''    By Martin Roesch & The Snort Team: http://www.snort.org/contact#team
           Copyright (C) 2014-2018 Cisco and/or its affiliates. All rights reserved.
           Copyright (C) 1998-2013 Sourcefire, Inc., et al.
           Using libpcap version 1.8.1
           Using PCRE version: 8.42 2018-03-20
           Using ZLIB version: 1.2.11

           Rules Engine: SF_SNORT_DETECTION_ENGINE  Version 3.0  <Build 1>
           Preprocessor Object: SF_FTPTELNET  Version 1.2  <Build 13>
           Preprocessor Object: SF_POP  Version 1.0  <Build 1>
           Preprocessor Object: SF_MODBUS  Version 1.1  <Build 1>
           Preprocessor Object: SF_DNP3  Version 1.1  <Build 1>
           Preprocessor Object: SF_GTP  Version 1.1  <Build 1>
           Preprocessor Object: SF_SSH  Version 1.1  <Build 3>
           Preprocessor Object: SF_SIP  Version 1.1  <Build 1>
           Preprocessor Object: SF_SDF  Version 1.1  <Build 1>
           Preprocessor Object: SF_DNS  Version 1.1  <Build 4>
           Preprocessor Object: SF_DCERPC2  Version 1.0  <Build 3>
           Preprocessor Object: SF_IMAP  Version 1.0  <Build 1>
           Preprocessor Object: SF_SMTP  Version 1.1  <Build 9>
           Preprocessor Object: appid  Version 1.1  <Build 5>
           Preprocessor Object: SF_SSLPP  Version 1.1  <Build 4>
           Preprocessor Object: SF_REPUTATION  Version 1.1  <Build 1>
Commencing packet processing (pid=2112)
03/27-02:42:40.254018  [**] [701:801:1] NMAP scan FIN [**] [Classification: Attempted
 Information Leak] [Priority: 2] {TCP} 192.168.43.224:42539 -> 192.168.43.88:22
```

Figure 9. Snort Testing

Figure 9. Snort Testing is a display of Snort Command Line Interface (CLI). The snort display usually in the form of CLI, but there are many third parties application that support Graphic User Interface (GUI). Testing snort can be done by input command on the terminal to start the snort program and snort will commencing packet data from network interface that snort listening from. If that packet match to snort rule, snort will give notification on terminal like shown on Figure 9.

4.3.2 Penetration Testing

Penetration on snort IDS is a stage in testing the network security monitoring system to find out whether the main application is running well or not.

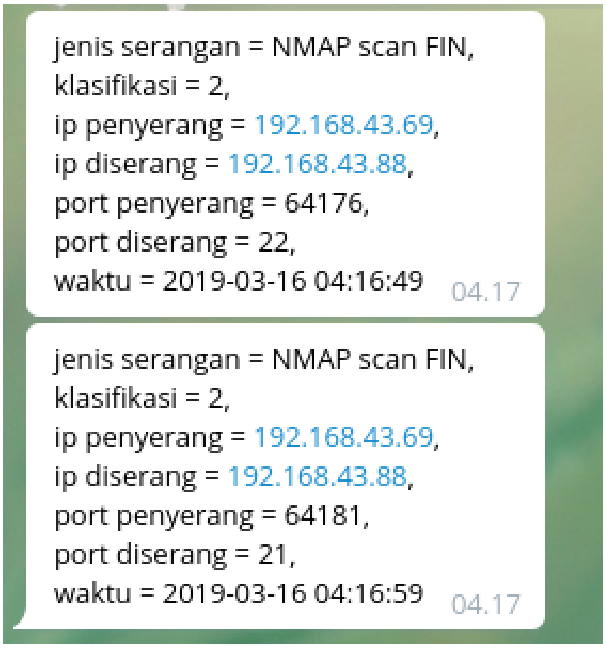


```
root@kali:~# nmap -sF 192.168.43.88
Starting Nmap 7.70 ( https://nmap.org ) at 2019-03-16 01:17 UTC
Nmap scan report for ghost (192.168.43.88)
Host is up (0.012s latency).
Not shown: 997 closed ports
PORT   STATE         SERVICE
21/tcp open|filtered ftp
22/tcp open|filtered ssh
80/tcp open|filtered http
MAC Address: BC:85:56:CE:78:5B (Hon Hai Precision Ind.)

Nmap done: 1 IP address (1 host up) scanned in 3.19 seconds
```

Figure 10. Penetration test using Nmap portscan

Figure 10. shown penetration test using Nmap portscan as a types of attacking from intruder computer with kali linux operation system. The intruder type nmap command and the victim IP address. The tool will start and shown what port are opened from the computer victim.

Penetration test using nmap portscan successfully launched and the bot telegram will send an information about malicious packet data who comes from nmap portscan attack like shown on Figure 11.



jenis serangan = NMAP scan FIN,
klasifikasi = 2,
ip penyerang = 192.168.43.69,
ip diserang = 192.168.43.88,
port penyerang = 64176,
port diserang = 22,
waktu = 2019-03-16 04:16:49 04.17

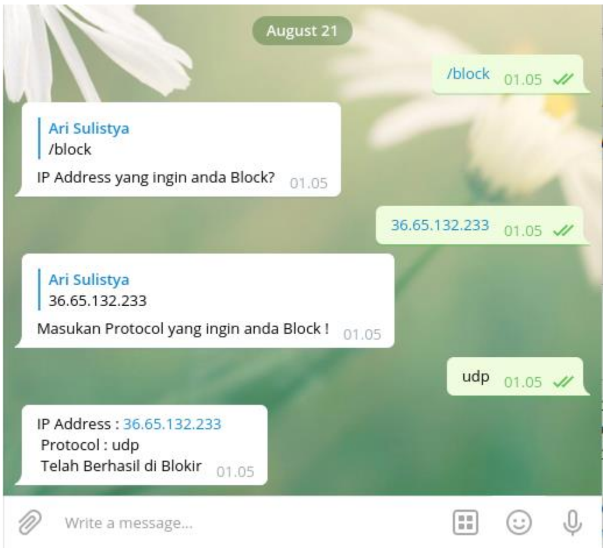
jenis serangan = NMAP scan FIN,
klasifikasi = 2,
ip penyerang = 192.168.43.69,
ip diserang = 192.168.43.88,
port penyerang = 64181,
port diserang = 21,
waktu = 2019-03-16 04:16:59 04.17

Figure 11. Telegram Information Messages

Figure 11. shown that Snort bot Telegram has receive information about malicious packet data from nmap portscan attacking. The inftomation that contains on this messages are The types of attacking, classification, Intruder IP address, victim IP address, intruder port, victim port, and time of attacking.

4.3.3 *Prevention Testing*

Prenvention system is a form of network security that works to prevent identified threats. Intrusion prevention systems continuously monitor your network, looking for possible malicious incidents and capturing information about them. This prevention system manually controlled by system administrator to block suspicious IP address.


August 21

/block 01.05

Ari Sulistya
/block
IP Address yang ingin anda Block? 01.05

36.65.132.233 01.05

Ari Sulistya
36.65.132.233
Masukan Protocol yang ingin anda Block ! 01.05

udp 01.05

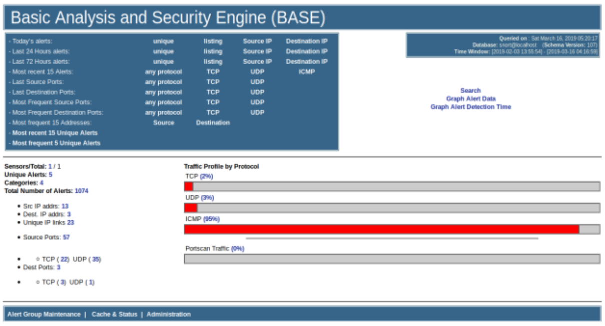
IP Address : 36.65.132.233
Protocol : udp
Telah Berhasil di Blokir 01.05

Figure 11. Prevention System

Figure 11. shown that bot Telegram can be use for prevention system. This features combine linux iptables to block IP address from suspicious client. System administrator just have to type block on this bot telegram and their suspicious IP address to blocking them from the system.

4.3.4 *Snort Web Based Interface*

Snort has many third parties web based interface, BASE is one of many web based interface. Basic Analysis and Security Engine (BASE) is to see all snort log without open the terminal.


Basic Analysis and Security Engine (BASE)

- Today's alerts:	unique	listing	Source IP	Destination IP
- Last 24 Hours alerts:	unique	listing	Source IP	Destination IP
- Last 72 Hours alerts:	unique	listing	Source IP	Destination IP
- Most recent 15 Alerts:	any protocol	TCP	UDP	ICMP
- Last Source Ports:	any protocol	TCP	UDP	
- Last Destination Ports:	any protocol	TCP	UDP	
- Most Frequent Source Ports:	any protocol	TCP	UDP	
- Most Frequent Destination Ports:	any protocol	TCP	UDP	
- Most frequent 15 Addresses:	Source	Destination		
- Most recent 15 Unique Alerts				
- Most frequent 5 Unique Alerts				

Queried on : Sat March 16, 2019 05:20:17
Database: snort@localhost (Schema Version: 107)
Time Window: [2019-02-03 13:55:54] - [2019-03-16 04:16:59]

Search
Graph Alert Data
Graph Alert Detection Time

Sensors/Total: 1 / 1
Unique Alerts: 5
Categories: 4
Total Number of Alerts: 1074

- Src IP addrs: 13
- Dest. IP addrs: 3
- Unique IP links 23
- Source Ports: 57
- ○ TCP (22) UDP (35)
- Dest Ports: 3
- ○ TCP (3) UDP (1)

Traffic Profile by Protocol
TCP (2%)
UDP (3%)
ICMP (95%)
Portscan Traffic (0%)

Alert Group Maintenance | Cache & Status | Administration

Figure 12. Snort Web Based Interface

Figure 12 shown BASE (Basic Analysis and Security Engine) this features analytics data of malicious packet data that snort detected. BASE features many analytics view such form protocol attack, even a graphic of computer system that detected suspicious packet data.

5. CONCLUSION

Snort Intrusion Detection System from testing of two types of penetration during 10 times of test successfully detected the suspicious packet data. There is delay from sending a messages from Telegram with value 4,05 seconds. Prevention system of this monitoring features can block IP address specified to port, protocol, and time during the block.

REFERENCES

[1] Ammad Uddin, Laiq Hasan, PhD, “Design and Analysis of Real-time Network Intrusion Detection and Prevention System using Open Source Tools”, University of Engineering and Technology, Peshawar, Pakistan. 2016.

[2] Parningotan Panggabean, S.Kom., M.Kom, “Analisis Network Security Snort Menggunakan Metode Intrusion Detection System (IDS) Untuk Optimasi Keamanan Jaringan Komputer”, Program Studi Sistem Informasi, STMIK GICI. 2018.

[3] Mukesh Sharma, Akhil Kaushik, Amit Sangwan, “Performance Analysis of Real Time Intrusion Detection and Prevention System using Snort”, The Technological Institute of Textile and Science,Bhiwani-127021, Haryana – India. 2012.

[4] Hargyo Tri Nugroho, Bagas Adi Wicaksono, “Utilizing Instant Messaging for Real-Time Notification and Information Retrieval of Snort Intrusion Detection System”, Department of Computer Engineering, Faculty of ICT, Universitas Multimedia Nusantara. 2013.

[5] Asep Fauzi Mutaqin, “Rancang Bangun Sistem Monitoring Keamanan Jaringan Prodi Teknik Informatika Melalui SMS Alert dengan Snort”. Universitas Tanjungpura. 2016.

[6] Bekti Maryuni Susanto, Agung Tri Guritno, “Implementasi Snort Ids Menggunakan Android Sebagai Media Notifikasi”,Politeknik Negeri Jember, 2017.Pp 1-3.

A Hybrid Data Loss Detection and Prevention Framework Using Snort Signature Based Detection System and Knowledge Based Anomaly Detection System

Henry Gekone Ondieki[1]
Jomo Kenyatta University of Agriculture and Technology
Masters Student
Nairobi, Kenya

Dr. Kennedy Ogada[2]
Jomo Kenyatta University of Agriculture and Technology
School of Computing.
Nairobi, Kenya

Prof. Wilson Cheruiyot[3]
Jomo Kenyatta University of Agriculture and Technology
Associate Professor of Computer Science.
Nairobi, Kenya

Abstract: In the modern day and era, data has become part and parcel of daily life and business. The concerns for data security have as a result of this, emerged as a major concern when seeking to prevent data leakages and data losses. The need to prevent unauthorized access to data has become a major factor affecting the survival of organizations today due to the consequences that could arise when data falls into the wrong hands. For instance, the level of credibility and trust-worthiness of various companies would be put into question wen sensitive data becomes accessed by unauthorized people. The existing traditional data security methods have not been enough preventative mechanisms to prevent loss and leakage of sensitive data. This calls for the development of a new and improved data security architecture creating the new data leakage prevention systems (IDSs/IPSs). Burgeoning research has seen new innovations and increased funding towards improvement of data security architecture. This study makes a contribution trough use of DLPs to propose hybrid data loss detection and prevention system. Signature based solutions provide accurate identification of the attacker and thus suitable for prevention, they cannot be used when unknown attacker or the attacker who uses different path attacks the system, also anomaly-based solutions can detect the unknown attacker but the false positive results are high thus limiting their allocation on systems. Due to this, in this thesis we propose a hybrid system which combines both the signature based and anomaly-based solutions which enables the detection and prevention of data loss.

Keywords: Hybrid Data Loss Detection, Snort Signature Based Detection System, Knowledge Based Anomaly Detection System

1. INTRODUCTION

In the current digital world, large amounts of data are being processed at every single moment due to the combination of various aspects of technology that include cell phones, internet and cloud computing which are part of business and daily life [1]. The amount of data that is processed on a daily basis is 2.5 quintillion bytes and it is on a steady growth on a steady growth because of an increasing data demand [2]. A major challenge facing the data industry is data loss and the need to protect information which is a key responsibility of data companies. This has seen the invention of methods to protect data from leakage. This led to the creation of Data Leakage Prevention systems (DLPs) which are meant to protect against data leakage and reduce the damage caused by data leakage [3].

DLP solutions ensure they carry out duties of prevention and to detect any attempts of unauthorized access to obtain sensitive data. Potential data breaches are prevented using the DLP in a way that is timely, which reduces the possible effects it could have created [4]. Signature-based IDS refers to the detection of attacks by looking for specific patterns, such as byte sequences in network traffic, or known malicious instruction sequences used by malware. The terminology is generated by anti- virus software, which refers to these detected patterns as signatures. Even though signature-based IDS can easily detect known attacks, it is impossible to detect new attacks, for which no pattern is available. This technique automatically contains the signature to detect an intruder [5].

The anomaly-based intrusion detection system refers to a method used in the detection of computer and network intrusions, and possible misuse through having a system that monitors activity and categorizing it to be either anomalous or normal. The categorization follows certain rules instead of signatures or patterns, and ascertains activity that is malicious and which do not appear to be normal operations of the system. Conversely, systems that are signature based are able to only detect types of attacks which already have previous signatures [5]. Knowledge based detection method is one of the Anomaly based IDS. In Knowledge based detection, knowledge is gathered on the attacks on data, and this knowledge is then used to detect any attacks or system vulnerabilities. When a system lacks knowledge about a particular attack, then it is not capable of identifying an attack. This means that the model requires a significant amount of knowledge on several attacks [6].

2. PROBLEM STATEMENT

Over the recent years, the challenge of ensuring data security has been acknowledged, and several methods have been developed to address the problem. Data leakage, corruption and loss are the main goals of attackers, and thus most of the methods that have been proposed aim at addressing this problem. Examined literature shows concerted efforts of data loss prevention systems.

A generic model of malicious behavior, distinguishing motives, actions and associated observables is in existence[7].

The study developed several prototypes that provided early warning of malicious activity including use of network traffic profiling, honey tokens and knowledge-based algorithms for data fusion and structure analysis. [8] proposed a hybrid intrusion detection system that detected and prevented malicious activity on a network. The method proved effective on evaluation for preventing data loss on a network and in dealing with considered attacks on data. [9] proposed a hybrid intrusion detection system in which a normal profile at activity intervals is first detected and then subsequently used to detect any anomalies in behavior nodes. Through considering anomaly types that detected, intrusion is then detected using predefined expert knowledge rules.

Existing data loss prevention methods continue to reveal loopholes that have been used by attackers who end up gaining unauthorized access to data. There is a need to develop and improve on existing data loss prevention methods. More secure data loss prevention methods will simply improve security architecture for data which will ensure accuracy, completeness, veracity and reliability of data. This study develops a hybrid data security method to protect against data loss by combining anomaly-based intrusion detection systems and signature-based intrusion detection system.

3. MAIN RESEARCH OBJECTIVE

The main objective of this study is to design and propose a hybrid data loss prevention system using a Snort Signature Based IDS and an Anomaly Knowledge Based Detection IDS.

3.1 Specific objectives

The specific objectives of this research are:

i. To study how the existing data loss prevention systems, work.
ii. To design a hybrid data loss prevention system using Snort Based Signature IDS and anomaly Knowledge based detection IDS.
iii. To evaluate the designed hybrid data loss prevention system.

The main aim of the hybrid data loss prevention system is to provide better security that prevents data loss in addition to providing a more secure hybrid data security system. An intrusion detection system (IDS) normally analyzes data from a network and detects any actions that are malicious and behaviors that may compromise the security of data. When activities that are malicious are detected, an alarm is raised. A hybrid data security combines the knowledge based detection anomaly and Snort signature based IDS to provide a more secure data security method which is the primary objective of this study.

3.2 Research questions

i) What are the existing data loss prevention methods and their shortcomings?
ii) How to design a hybrid data loss prevention system using anomaly Knowledge based detection model and Snort signature-based IDS?
iii) How will this developed hybrid system be evaluated?

4. LITERATURE REVIEW

4.1 Snort Based IDS

Snort is an open and free system for prevention of intrusions created by Martin Roesch in 1998.

4.1.1 Components of Snort

Snort is categorized into different components that all work in collaboration with each other to detect specific attacks and for the generation of output from the detection system a format that is pre-specified [10].

The packet decoder selects packets from different network interfaces and prepares them to be pre-processed or sent to the detection engine. Pre-processors are components or plug-ins that are to be used with Snort for modifying or arranging packets before performance of any operation by the detection engine in order to establish if any packets have been exploited by an intruder. The detection engine is responsible for detecting any intrusion that may exist in a packet based on rules of Snort [10].

4.1.2 Anomaly Based IDS

Anomaly detection methods are very important in intrusion detection systems because an intrusion activity is different from the normal activity of the system. Anomaly based IDS uses a reference of a pattern of normal system activity that has been learned or baseline in order to establish active intrusion attempts. Any behavior that fall outside the accepted model of behavior or pre-defined pattern creates an anomaly [6.]

4.1.3 Knowledge Based Detection

Knowledge based detection obtains knowledge about attacks and this knowledge is then used to detect any system vulnerabilities or attacks. They are useful in recognizing transitions that occur when an intruder penetrates the security system. Expert systems contain sets of facts, rules, and inferences methods. Each event that occurs in the system is translated into corresponding rues and facts and inference methods are able to generate conclusions from the existing rules and facts. Signature analysis contains similar knowledge acquisition approach as in an expert system, but the way of knowledge acquired is different. The exact evidence of every attack is available in the audit trail and this information is consolidated as sematic description of attack [6]

4.2 Data Leakage Prevention (DLP)

Data Leakage Prevention (DLP) refers to one of the specialized philosophies and arrangements that are used to protect information that is delicate from being accessed by unauthorized people who are either inside or outside the organization [11.] DLP is a method used to conceal the secrecy of information being gotten to by unapproved client [12]. DLP arrangements address information leakages through three different categories by using specific types of technology arrangements [13; 14; 15].

4.3 Data security threats and vulnerability

To begin with, issues of data security and protection have grown as a result of volume, speed, and assortments. For instance, large scale infrastructure used for cloud computing, rich sources of data and configurations, nature of information spilling, and large volume between cloud relocation [16]. Further, usage of these large-scale infrastructures for cloud computing are spread around the globe with different types of software have seen an increase in system attacks and thus traditional systems for security are inadequate. Further, better and more improved technologies that are able to quickly respond to the growing demands of streaming data across several centers of data [16].

4.4 Data loss prevention Techniques

Techniques for data loss prevention can be grouped into non-sensitive data and sensitive data and thus techniques or detective purposes are categorized into two main areas i.e. content-based analysis techniques and context –based techniques which are discussed under [17;18].

4.4.1 Context analysis technique:

This technique works through considering metadata (format, source, size, timing, and destination) that is often linked to the actual confidential data without emphasizing on how sensitive the data is. The Dkey contextual features such as size, source, timing, and destination would be examined. The system then compares these features with certain patterns of transactions or pre-defined policies. This technique is sometimes combined with content-based analysis techniques for it to be effective.

4.4.2 Data fingerprinting:

This is the most widely recognized strategy which is utilized to distinguish information spillage. In many DLPSs, an entire record can be hashed utilizing ordinary hash capacities, for example, MD5 and SHA1, where the hash values of each single delicate archive are stored in nearby machines or databases. These DLPs are able to have 100% accuracy in identifying whether a record changed by using any and all available means. Records that are secret are prone to being changed; DLPs may be inadequate because hash esteem cannot defend itself from change. This then means conventional fingerprinting methods are infective to prevent significant changes to the data. This can be solved using more advanced fingerprinting methods for instance Rabin's randomized fingerprinting and fuzzy fingerprinting.

4.4.3 Regular expression:

Most DLPs use this famous technique. These DLPs are made of set of terms or characters that are utilized to frame location designs. These examples will be utilized to match and think about arrangement of information strings numerically. This procedure is for the most part utilized as a part of web crawlers and content preparing to approve, extricate and supplant information. In any case, as far as data security, consistent articulation is utilized generally in information examination for vindictive codes or secret information.

4.4.4 Statistical examination:

This strategy can encourage certain devices, for example, machine learning order and data recovery term weighting. Most part of them depends on the terms and n-grams recurrence inside arrangement of archives. The disadvantage of consistent articulation and information fingerprinting were tackled by N-gram measurable examination procedure. A term basically implies a word, while a n-gram may be bits or a word E.g. unigram (one character), bigram (two characters) and trigram (three characters).

4.5 Preventive method

Policy and Access Rights: Prevention of possible data leaks using strict access controls. Some organizations have policies that restrict use of CDs and USB.

Virtualization and Isolation: these are used to protect data that is sensitive from through h ensuring the creation of virtual environments when accessing data that is sensitive. Access that is allowed will be the only one permitted.

Cryptographic Approaches: This is a method used to hide data that is sensitive from being accessed by users who are not authorized by using algorithms and cryptographic tools. [19] used Attribute Based Encryption (ABE) algorithm a prevention method for data leakage which allowed sensitive data to be preserved. This is a preventative method. The system worked by keeping sensitive data locked and only allow users that are authorized to access it. The idea is that reliance on detective approaches can result in data leakage. Encryption prevents such data losses from occurring.

Quantifying and Limiting: Security administrators use this method to pretend to be system attackers and block all possible loopholes that lead to data that is sensitive by making attacks on their own systems. This approach is applicable for both detection and prevention methods.

4.6 Detective Method:

Data Identification: Refers to the way in which data that is sensitive is detected depending on the previous knowledge of content and some techniques such as data fingerprints among other types of matching.

Social and Behavioral: analysis of Patterns and behaviors on Social network can enable detection of irregularity and raise alarms to enable security administrators take action.

Data Mining / Text Clustering: Data mining areas have capabilities to perform advanced undertakings, for example, inconsistency location, bunching and order by removing information designs from vast datasets. Identification of Information mining by using machine realizing that has algorithms to establish patterns that are complex and enable better decision making. Clustering of texts is related to retrieval of information that is essential in DLPs.

5.0 Methodology

In this study experimental research design is used. Publicly available datasets were used in the literature for testing Intrusion Detection Systems. Such datasets served as a benchmark for the various parameters of an IDS like false positives, false negatives and detection rates. This also served the purpose of analyzing such parameters in relation to other existing IDS systems. Parameters like true and false positives rates, true and false negatives rates are the most important parameters of any IDS. These parameters are a measure of the effectiveness of the detection mechanism of an IDS. For the given dataset, the results of these parameters were analyzed for different thresholds and r-values. ROC curve analysis was used to test the effectiveness of the IDS systems.

6.0 Findings and Discussions

The developed system was tested against the dataset and parameter values like false positives, false negatives and detection rates were calculated. ROC curves were also plotted for the obtained results. Testing was repeated with different values of n, r, and anomaly threshold values and the corresponding rates were calculated.

6.1 Processing data

Some pre-processing had to be done on the dataset before testing it on the hybrid data loss detection and prevention system. Since the negative selection module accepts only self-traffic as an input for training, a Python script was written to retrieve only the normal traffic from the labeled training data.

Also, since the self-data has to have strings that are of the same length, the records were padded with the character ‘0’ so that all the records were the length of the largest string in the file. Similarly, a different Python script was used for calculating false positive, false negative, and detection rates from the logs of the developed hybrid data loss detection and prevention system.

6.2 Detection and false alarm rates

False positive and detection rates were calculated by running the test data dataset. These rates were obtained for different n, r and threshold values. Since the length of strings in the test dataset was 152 (including the added padding), the n values tested were between 70 and 120. Higher n values only resulted in poorer detection rates and insufficient heap memory. As stated earlier, r values were tested between 40% and 80% of the given n value. Anomaly threshold was varied between 10% and 100%. Highest detection rate obtained was 81.56% for n=100, r=40 and threshold = 20%. But the corresponding false positive rate was also high at 39.52%.

By varying the above mentioned parameters, the false alarm and detection rates were calculated. Table 4.1 shows the different average rates for given n values. It can be seen than an n value of 100 yields the best average detection rate of 64.90%. But this also resulted in a moderate average false positive rate of 31.37%. As n value was increased above 100, the average detection rate started to decline. This may be attributed to the increased noise being included for generation of detectors. Also with higher n values, the system ran out of heap space and crashed abruptly. This can be attributed to huge storage overhead associated with generation of large detectors. The percentage of test cases crashed for given n values can be seen in the Figure 4.1.

6.3 ROC Curves

ROC curves are plots of detection rates against false positive rates. Such curves indicate the operating region of a system and also lets the network administrator decide on the operating region that is suitable for a given network environment. As mentioned earlier, the points (0,0) and (1,1) occur at the worst operating conditions of any hybrid data loss detection and prevention system. Figure 4.2 shows the ROC curve for the tested dataset with these two points. This curve was plotted using about 565 tested cases, obtained by varying n, r and threshold values.

Table 4.1: Comparison of average rates for given n values

	Average False	Average	Average False
n-value			
	Positive rate (%)	Detection Rate (%)	Negative Rate (%)
70	28.81126149	61.32805454	38.67195
75	31.37546036	59.14728688	40.85271
80	32.26197881	58.73763987	41.26236
85	32.56920446	58.97770202	41.0223
90	32.59815859	59.43509852	40.5649
95	32.25811202	59.92991625	40.07008
100	31.37725954	64.9015537	35.09845

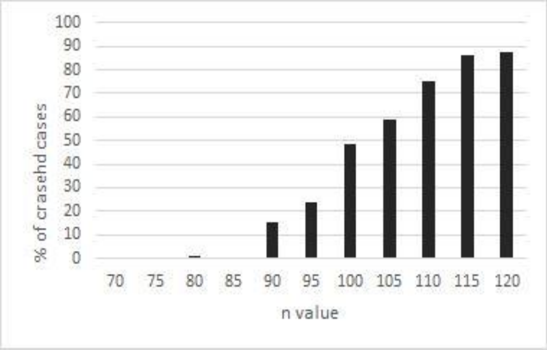


Figure 4.1.: Percentage of crashed test cases Vs. n value

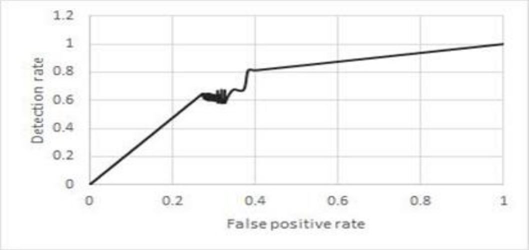


Figure 4.2.: ROC curve including points (0,0) and (1,1)

Figure 4.3 shows the actual ROC curve without the points (0,0) and (1,1). This curve is generally convex and does not have abrupt drops in slope. The region from false positive rate of 0.27 to slightly after 0.32 can be considered as ideal operating region for this system. With the increase in false positives beyond 0.27, there is no abrupt drop in detection rate, indicating a wide operating region with decent false positive and detection rate. There is a slight concave region in the curve where false positive rate is 0.37. This is a sign of poor detection mechanism in that region, but since it is outside the ideal operating region of the system where the false positive rate is already higher, this is not a huge compromise on the effectiveness.

4.4 Effect of r-value on detection and false positive rates

The r values tested ranged from 40% to 80% of the value of n (calculated from (r/n)*100). The false positive and detection rates against various r value percentages can be seen in the Figure 4.4. Individual detection rate (at 81.56%) and average detection rate (at 62.68%) were the maximum when the r value was 40% of n value. But the average false positive rate was also slightly higher at 30.50%. It is to be noted that these are values are specific to the dataset and may not always be true for another dataset. Ideal r and n values for a different dataset can be deduced by similar testing.

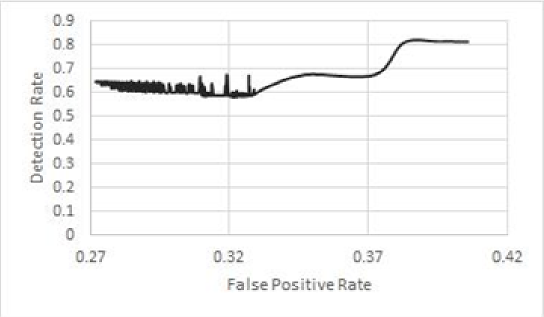


Figure 4.3.: ROC curve

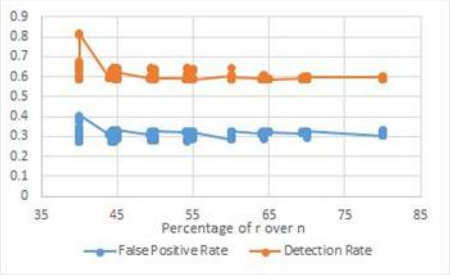


Figure 4.4.: False positive and detection rates against percentage of r values over n

4.5 Comparison of effectiveness with other IDS systems

Figure 4.5 shows the comparison of false positives and detection for different r values and constant n and threshold values. This graph shows the direct effect of r value on the detection rate. The n value was set at 75 and threshold value was set at 70%. It could be observed that as the r value increases, detection rate decreases and the false positive rate increases. As already seen in Figure 4.1, higher n values resulted in heap spa running out and the program crashing. The insufficient heap space issue further increased with higher r values. It can be concluded that for the dataset, the effectiveness and performance of the system decreases beyond an r percentage of 40%.

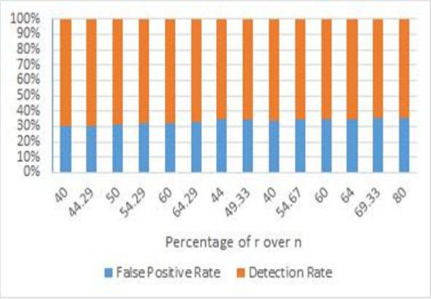


Figure 4.5.: False positive and detection rates against percentage of r values over n (constant n and threshold values)

The Figure 4.6 shows the ROC curves of the developed hybrid data loss detection and prevention system along with the other prominent systems like Snort (Alder et al., 2007), Bro (Paxson, 1999), Hybrid Intrusion Detection System (HIDS) proposed in Hwang, Cai, Chen, and Qin (2007). ROC curves for these three IDS systems were already compared in Hwang et al. (2007). HIDS is a combination of anomaly-based and signature-based detection mechanisms and would serve as a proper comparison. HIDS having a combination of anomaly based and signature based detection mechanisms, performs better overall compared the other three IDS systems. The operating range of HIDS has a detection rate around 30% better than Snort and around 38% better than Bro. In comparison with the developed IDS, HIDS has almost an identical detection rate range in the operating region before a false positive rate of 0.32

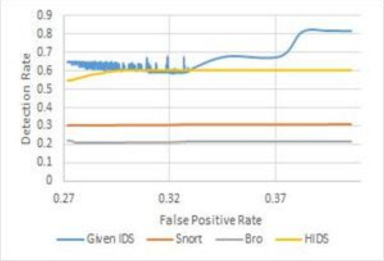


Figure 4.6.: ROC curves - comparison against other IDS systems.

4.6 Effectiveness against zero-day attacks

It is to be noted that the developed IDS uses only self-traffic for training and not non-self-traffic. Thus, in the view of the IDS, all the attacks are unknown (like zero day attacks). This makes the detection rates and false positive rates mentioned above equally applicable to zero days' attacks as it is to any other attack. Even so, the dataset has only 568 (out of the total of 22,544 records) of the non-self-instances in the test dataset carried over from the training dataset. That is, only 2.43% of the entries in the test dataset is comprised of previously known attacks.

5.0 SYSTEM ANALYSIS AND DEVELOPMENT

To overcome the shortcomings of existing intrusion detection systems, a multi-layer model is provided (Figure 5.1) which consists of three processing layers: 1) Packet Analysis; 2) Intrusion Detection; and 3) Security Information and Event Management (SIEM).

5.1 Packet Analysis

Being responsible for all the preprocessing tasks required for the intrusion detection, Packet analysis layer contains two important modules, namely flow analyzer and traffic classification.

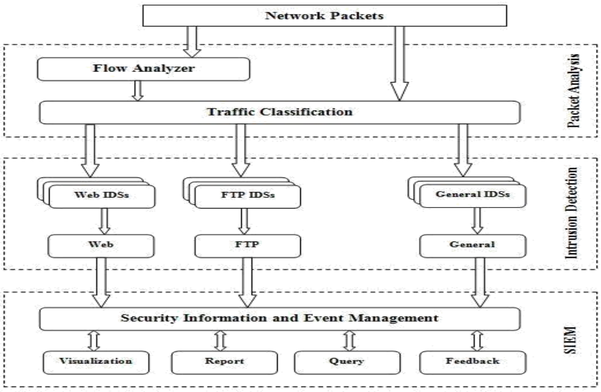


Figure 5.1: The proposed framework

5.1.2 Flow Analyzer

In order to keep up with the high speed of gigabit links, current network monitoring and management systems use network flow data (e.g., netflow, sflow, ipfix) as their information sources. Network flows are a group of net-work packets belonging to the same connection. Apparently, these packets have a lot of information in common (e.g., source IP, source port, destination IP, destination port, protocol, application), which can be stored only once using the flow concept. Furthermore, in order to deal with a huge amount of payload information, usually only the first few bytes of each flow (e.g., 512 bytes), which is more informative for the analysis, will be stored by the monitoring devices.

5.1.3 Traffic Classification

As network applications are getting more diverse and complex, the idea of using special-purpose intrusion detection systems in the application layer be-comes more popular. Focusing on a small subset of network applications will have the advantage of designing more specific signatures, which results in a better detection rate and a lower false positive rate. In addition, as illustrated in Figure 5.1, application-based IDSs can be applied in parallel which is of high importance in dealing with large networks with millions of packets per second.

5.2 Intrusion Detection

As indicated in Figure 5.1, the proposed intrusion detection module consists of several application-based intrusion detection systems components. Sharing a similar architecture and detection mechanism, each component is specifically designed for a special type of application such as Web, FTP, Mail, etc. There is also a component designed for applications with no specialized intrusion detection system and applications that are not detected with the traffic classifier.

5.2.1 Security Information and Event Management

Although a lot of efforts have been done to decrease the number of false alarms generated by intrusion detection systems, we believe that having an IDS with no false alarm is almost impossible due to the dynamic nature of computer networks. However, we can minimize these false alarms by gathering and processing different types of information from various sources such as intrusion detection systems, anti-viruses, operating systems logs, application level logs, among others.

5.3 Traffic Classification Module

Accurate classification of network traffic has received a lot of attention due to its important role in many subjects such as network planning, QoS provisioning, class of service mapping, to name a few. Traditionally, traffic classification relied to a large extent on the association of a particular port with a specific protocol. Such a port number based traffic classification approach has been proved to be ineffective due to: 1) the constant emergence of new peer-to-peer networking applications that IANA does not define the corresponding port numbers; 2) the dynamic port number assignment for some applications (e.g. FTP); and 3) the encapsulation of different services into a same application (e.g., chat or steaming can be encapsulated into the same HTTP protocol). To overcome this issue, there have been recently significant contributions towards traffic classification. The most currently successful approach is to inspect the content of payloads and look for the deterministic character strings for modeling the applications. For most applications, their initial protocol handshake steps are usually different and thus can be used for classification.

5.3.1 Weighted Unigram Model

N-grams are a language-independent means of gauging topical similarity in text documents. Traditionally, the n-grams technique refers to passing a sliding window of *n* characters over a text document and counting the occurrence of each n-gram. This method is widely employed in many language analysis tasks as well as network security. Applying the same idea on network packets, one can consider unigram (1-gram) of a network packet as a sequence of ASCII characters ranging from 0 to 255. This way similar packets can be identified using the frequencies of distinct ASCII.

5.3.2 Problem Formulation

In this section, we formally describe how the network application discovery problem can be performed through the combination of genetic algorithms and decision trees. Essentially, we formulate the network application discovery problem as a classification problem, i.e., given the values for a specific set of features extracted from the network flows, we identify the possible application that has generated this payload using a statistical machine learning technique (decision tree).

5.4 Intrusion Detection Module

Traditionally, intrusion detection techniques are classified into two categories: misuse (signature-based) detection and anomaly detection. Misuse detection is based on the assumption that a large number of cyber-attacks leave a set of signatures in the stream of network packets or in audit trails, and thus attacks are detectable if these signatures can be identified by analyzing the audit trails or network traffic behavior. However, misuse detection is strictly limited to the known attacks and detecting new attacks is one of the biggest challenges faced by misuse detection.

5.4.1 Anomaly-based Detector

As the first step to have an effective anomaly detector, we should extract robust network features that have the potential to discriminate anomalous behavior from normal network activities. Since most current network intrusion detection systems use network flow data (e.g. netflow, sflow, ipfix) as their information sources, we focus on features generated based on these flows.

5.4.2 Signature-based Detector

As our first signature-based detector we chose Snort because of its popularity and availability to researchers. However, our proposed hybrid detection scheme is completely independent from Snort, and any other signature-based detector can be used instead. As mentioned earlier, our anomaly-based detector works on flows. However, Snort is designed to work on packets.

6.0 CONCLUSION AND RECOMMENDATION

A system is proposed which has an adaptive hybrid data loss detection and prevention intrusion detection system to overcome the main shortcomings of the existing IDSs.

With regard to the hybrid intrusion detection, we have identified two main is-sues that highly affects the performance of the system. First, anomaly-based methods cannot achieve an outstanding performance without a comprehensive labeled and up-to-date training set with all different attack types, which is very costly and time-consuming to create if not impossible. Second, efficient and effective fusion of several detection technologies becomes a big challenge for building an operational hybrid intrusion detection system. To solve the first issue, we have proposed applying the idea of adaptive learning. To meet this goal, we have defined learning time intervals, e.g. 1 day, at the end of which the anomaly-based detector will be trained by the two most recent training sets. These training sets are the flows that are labeled by the hybrid detector in the previous intervals.

7. REFERENCES

[1] Mahajan, P., Gaba, G., & Chauhan, N. S. (2016). Big Data Security. IITM Journal of Management and IT, 7(1), 89-94.

[2] Harish Kumar, M. & Menakadevi, T. (2017), A Review on Big Data Analytics in the field of Agriculture, International Journal of Latest Transactions in Engineering and Science, vol. 1, issue 4, pp. 0001-0010. Hevner, A. R., March, S. T., Park, J. & Ram, S. (2004), Design Science in Information Systems Research, MIS Quarterly, vol. 28, no. 1, pp. 75-105.

[3] Sagiroglu, S., and Sinanc, D. 2013. "Big data: A review," in Proceeding of the International Conference on Collaboration Technologies and Systems (CTS), pp. 42–47 (doi: 10.1109/CTS.2013.6567202).

[4] Fang, Z., & Li, P. (2014). The mechanism of "big data" impact on consumer behavior. American Journal of Industrial and Business Management, 4(1), 45-50.

[5] Tiwari, M., Kumar, R., Bharti, A. & Kishan, J. (2017). Intrusion Detection System. International Journal of Technical Research and Applications, 5(2):2320-8163. Retrieved from https://www.ijtra.com/view/intrusion-detection-system.pdf?paper=intrusion-detection-system.pdf

[6] Jose, S., Malathi, D., Reddy, B., Jayaseeli, D. (2018). A Survey on Anomaly Based Host Intrusion Detection System. Journal of Physics: Conf. Series 1000 (2018) 012049 doi :10.1088/1742-6596/1000/1/012049

[7] Maybury, M.T., Chase, P., Cheikes, B.A., Brackney, D., Matzner, S., Wood, B.J., Longstaff, T., Hetherington, T., Marin, J., Spitzner, L., Copeland, J.S., Lewandowski, S.M., & Haile, J. (2005). Analysis and Detection of Malicious Insiders.

[8] Granjal, J., & Pedroso, A. (2018). An intrusion detection and prevention framework for internet-integrated CoAP WSN. Security and Communication Networks, vol. 2018, Article ID 1753897, 14 pages, 2018.

[9] Desnitsky, Vasily & Kotenko, Igor & Nogin, S.. (2015). Detection of anomalies in data for monitoring of security components in the Internet of Things. 189-192. 10.1109/SCM.2015.7190452.

[10] Funke Olanrewaju, R., Ul Islam Khan, B., Rahman Najeeb, A., Afiza Ku Zahir, K., & Hussain, S. (2018). Snort-Based Smart and Swift Intrusion Detection System. Indian Journal Of Science And Technology, 11(4). doi:10.17485/ijst/2018/v11i4/120917

[11] Kale, A. V., Bajpayee, V. & Dubey, S. P. (2015), Analysis of Data Leakage Prevention Solutions, International Journal For Engineering Applications And Technology (IJFEAT), vol. 1, issue, 12, pp. 54- 57.

[12] Jain, M & Lenka, S. K. (2016), A Review on Data Leakage Prevention using Image Steganography, International Journal of Computer Science Engineering (IJCSE), vol. 5, no. 02, pp. 56-59.

[13] Tahboub, Radwan & Saleh, Yousef. (2015). Precaution Model for Data Leakage Prevention/Loss (DLP) Systems.

[14] S. W. Ahmad and G. R. Bamnote, "Data Leakage Detection and Data Prevention using Algorithm," International Journal of Computer Science and Application, vol. 6, pp. 394-399, 2013.

[15] Peneti, S. & Rani, B. P. (2015a), Data Leakage Detection and Prevention Methods: Survey. Discovery, vol. 43, no. 198, pp. 95-100.

[16] Shirudkar, K. & Motwani, D. (2015), Big-Data Security. International Journal of Advanced Research in Computer Science and Software Engineering, vol. 5, issue 3, pp. 1100-1109.

[17] Alneyadi, S., Sithirasenan, E. & Muthukkumarasamy, V. (2016), A survey on data leakage prevention systems, Journal of Network and Computer Applications, vol. 62, issue C, pp. 137-152.

[18] Alneyadi, S., Sithirasenan, E. and Muthukkumarasamy, V. (2015), Detecting Data Semantic: A Data Leakage Prevention Approach, In the Proceedings of the 2015 IEEE Trustcom/BigDataSE/ISPA, August 20 - 22, IEEE Computer Society Washington DC, USA, vol. 1, pp. 910-917.

[19] Margathavalli, P., Manjula, R., Pramila, V., Priya, R. & Abirami, P. (2016), Preserving Sensitive Data by Data Leakage Prevention Using Attribute Based Encryption Algorithm, International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE), vol. 21, issue 3, pp. 705-711.

International Journal of Computer Applications Technology and Research
Volume 9–Issue 02, 72-75, 2019, ISSN:-2319–8656

Kenya Road Accidents Cause Classification Using Bayesian Networks

Raphael Ngigi Wanjiku
School of Computing and Information Technology
Jomo Kenyatta University of Agriculture and Technology
Nairobi, Kenya

Abstract: In Kenya, the number of fatalities from road accidents rise year after year due to various causes. However, these numbers differ year after year and it is very difficult to identify the causation making analysis and management of anti-accident public campaigns difficult. With the use of Bayesian networks, the causal analysis can be probabilistically estimated giving a better analysis and therefore better measures in addressing the underlying causes. This paper utilises data from the Kenya National Transport Safety Authority website which is pre-processed and prepared for use in a Bayesian network model. Thereafter a Bayesian network model is built using 70% of the dataset as the training data and 30% as testing data. The model is developed with the aid of the Weka software utilising a sample of 120 instances from the prepared data with 401 instances. Furthermore, to validate the model, a Naïve Bayes model is developed with the same dataset. The Bayesian network model results in 69.125% accuracy which is lower compared to those given by the naïve Bayes model with 72.5% accuracy possibly due to the fact that Naïve Bayes algorithm performs well even with small amounts of data. Also, from the results, the model identifies that most of the accidents are driver related with 63.8% on the Bayesian network and 78.2% on the Naïve Bayes model and therefore more need to be done in addressing the driver causes. However, more variables need to be introduced in the dataset by the transport agency.

Keywords: Bayesian network, National Transport Safety Authority, normalization, Naïve Bayes, Matatu.

1. INTRODUCTION

The number of people who die on the Kenyan roads is worrying with the year 2018 recording 2965 deaths [6] at the scene of crash. However, many other people die from subsequent effects of the accidents and therefore proper analysis of the causes could alleviate proper campaigns to help reduce the fatalities.

In Kenya, accidents are recorded by the National Transport and Safety Authority which keep daily and monthly reports. The accidents are tabulated in Excel spreadsheets with simple analysis of variations in the monthly reports. This data is then shared among the stake-holders, mostly the Kenya Police Service in ensuring motorists follow the Traffic Act [1]. Despite all the efforts being made, these accidents and especially 2019 have increased by 13% compared to the year 2018. Looking at the previously available data, these accident patterns vary over the years and alternative ways of addressing the causes through technology could be quite beneficial-Bayesian networks in artificial intelligence.

A Bayesian network refer to a probabilistic graphical model that effectively deals with various uncertainty problems. [2] Bayesian Networks are normally used to detect causal relationships for example over speeding and carelessness causing road accidents. They have been used in fault diagnosis [3], Customer satisfaction in public transport management [4], and prediction of criminal cases [5].

Bayesian networks aim at modeling conditional dependence showing causation among variables [7]. They are built from a probability distribution since they work with probabilities. The network is made of nodes and arcs connecting the nodes. Each arc has an arrow that point from a parent variable and the network utilises conditional probability. Given four nodes A, B, C and D as shown in the figure one below, A is considered a parent of A due to the direct causal relationship while D is C is independent of A since there is no direct connection between the two nodes. Each of the node for example B has a conditional probability distribution P (B | Parent(B)).

$$P(D) = P\left(Yi \middle| parent\,(Yi)\right) \qquad (1)$$

$$P(D) = \frac{P(parent\,(Yi)|Yi)P(Yi)}{P(parent\,(Yi))} \qquad (2)$$

meaning calculating the probability of B given the probability of A (since A is the parent of B).

A Bayesian network utilises a joint probability distribution and the conditional probability. A joint probability distribution of the variables [8] shown in figure 1 is P (A, B, C, D).

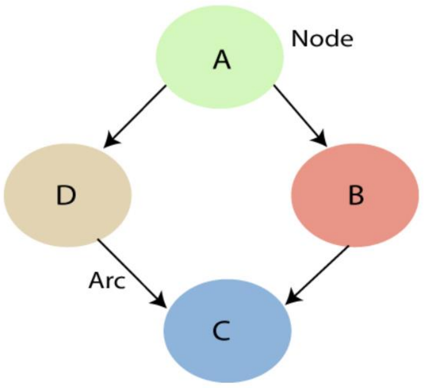


Figure 1. A classification-type decision tree output for this prediction domain.

2. RELATED WORK

There are several literatures works on the utilization of Bayesian networks in evaluating road safety. In a study by Mujalli [10], various Bayes classifiers were used to identify the factors that affect the severity of an accident using Jordan accidents data for the years 2009-2011.They used balanced databases which improved the performance of the classification to the original classification[11]. They were also able to identify other factors that lead to serious injuries or fatalities in accidents for example, the number of vehicles involved, speed limits and the type of accident.

In a different study conducted by De Ona and Mujalli [10], Bayesian networks were used to model accidents classification in Spain (moderate, severe and fatal) and the study concluded that the number of probability variables can be reduced and still maintain the accuracy of the network.

Other works conducted by Zou [2] similarly develop models that address the severity of accidents. Further work by Deublein [8] had shown how Bayesian networks can be used in the prediction of road accidents with case study of Austrian rural motorway network using multivariate regression analysis.

This paper primarily addresses the relationships that exist among the vehicles involved in accidents (car, bus, cycle, tuk-tuk, lorry, other vehicles), time of accident occurrence, gender casualty, pedestrians involved and the Kenyan county most likely to happen.

3. METHODOLOGY

3.1 Accidents database

The study used the NTSA data released every month. The objective was to determine the causes of the accidents and the other related variables. There were difficulties in obtaining complete records hence the 401 instances used do not exhaustively report all the accidents that have occurred with fatalities in Kenya. There were other errors noted in the recording of the data including double entries and mistyped data.

In the study, twenty (20) variables were used to determine the relationships that exist between the various classification of fatalities: Period parameters (year, time),Kenyan county, vehicle involved (car, bus, cycle, tuk-tuk, lorry, matatu, other vehicle, unknown vehicle), cause, gender(male, female), NTSA code, driver, victim involved (passenger, pedestrian, cyclist) and number of casualties.

3.2 Bayesian network modelling

Out of the 401 provided instances, the software selected a sample of 120 instances as the dataset and dividing it into two sets: training data (70%) and the rest (30%) as the test data. The modeling process was aided by the Weka software, a data mining tool developed by the University of Waikato, New Zealand.

The twenty (20) variables were further reduced to fifteen (15) variables and three classification classes were developed: driver, pedestrian and vehicle. The values of the variables were coded into discrete values for computational purposes as shown in the figure below.

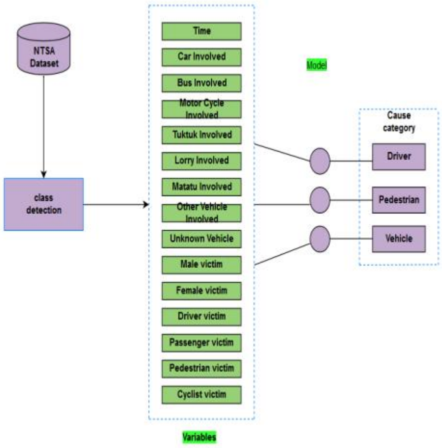


Figure 2. A Bayesian network model on the NTSA data.

The raw data was rescaled to make it suitable for modeling on the Weka software through normalization of the attribute values. Normalization refers to the process of changing the numeric columns in the dataset to use a common scale without distorting differences in the ranges of values or losing information [9]. In this case, all the attributes used have been rescaled to be in the range of 0 and 1.

4. EXPERIMENTS AND RESULTS ANALYSIS

A total of 120 sampled instances gave an accuracy of 69.17% representing 83 correctly classified instances and 30.83% representing 37 instances as wrongly classified which is shown in the confusion matrix in table 1.

Table 1. Confusion matrix on the cause classes

Cause class	Vehicle	Driver	Pedestrian
Vehicle	32	5	0
Driver	29	51	0
Pedestrian	2	1	0

From table 2 below, the precision of the model using the Bayesian net-work shows that it could not classify the accidents resulting from pedestrians' faults while giving a precision of 50.8% for the accidents resulting from vehicle related issues and 63.8% for the driver related accidents.

Table 2. Accuracy of the Bayesian network

Cause class	TP Rate	FP Rate	Precision	F-Measure	PRC Area
Vehicle	0.865	0.375	0.508	0.640	0.551
Driver	0.638	0.895	0.638	0.745	0.838
Pedestrian	0.000	0.000	0.000	0.000	0.025
Weighted Average	0.692	0.215	0.000	0.000	0.729

Table 3. Confusion matrix on the cause classes using the Naïve Bayes network

Cause class	Vehicle	Driver	Pedestrian
Vehicle	18	17	2
Driver	11	68	1
Pedestrian	0	2	1

Table 4. Accuracy of the Naïve Bayes network

Cause class	TP Rate	FP Rate	Precision	F-Measure	PRC Area
Vehicle	0.486	0.133	0.621	0.545	0.705
Driver	0.850	0.475	0.782	0.850	0.861
Pedestrian	0.333	0.026	0.333	0.286	0.140
Weighted Average	0.692	0.215	0.000	0.000	0.729

Table 5. Model variables means and standard deviations

Variable	Vehicle		Driver		Pedestrian	
	Mean	Std	Mean	Std	Mean	Std
Time	0.5874	0.2737	0.5403	0.2962	0.5865	0.3092
Car	0.4436	0.4968	0.4048	0.4908	0.2500	0.4330
Bus	0.0752	0.2637	0.1071	0.3093	0.2500	0.4330
Motor cycle	0.2556	0.4362	0.2341	0.4235	0.0000	0.1667
Tuk-tuk	0.0977	0.2970	0.0476	0.2130	0.0000	0.1667
Lorry	0.2707	0.4443	0.1905	0.3927	0.1875	0.3903
Matatu	0.1654	0.3716	0.1310	0.3373	0.1250	0.3307
Other vehicle	0.0752	0.2637	0.0317	0.1753	0.1250	0.3307
Unknown vehicle	0.0301	0.1708	0.2183	0.4131	0.1250	0.3307
Male	0.2462	0.1371	0.2054	0.1032	0.2188	0.0827
Female	0.0739	0.0186	0.0344	0.0706	0.0208	0.0551
Driver	0.2331	0.4228	0.0437	0.2043	0.0000	0.1667
Passenger	0.3910	0.4880	0.1548	0.3617	0.5000	0.5000
Pedestrian	0.1128	0.3163	0.5675	0.4954	0.5000	0.5000
Cyclist	0.3233	0.4677	0.2421	0.4283	0.0000	0.1667

From table 4, the Naïve Bayes network performed a better classification of test data giving a 33.3% for accidents caused by pedestrians and 78.2% for the driver caused accidents performing better compared to the Bayesian network which gave a 63.8% for the same sampled dataset.

The same data modeled using a Naïve Bayes gives an accuracy of 72.5% (87 instances) wrongly classifying 27.5% (33 instances).

5. CONCLUSION AND RECOMMENDATIONS FOR FUTURE WORKS

In this paper, a model showing causation of Kenyan accidents fatalities selected in the years 2017 to 2019.The methodology utilises a Bayesian model whose performance(precision) is compared with the naïve Bayes model.

From the results, a naïve Bayes gives a better precision for the classification giving a 72.5% compared to 69.125% given by the Bayesian network. This is possibly due to the fact that it requires less data compared to the Bayesian network with the selected sample of 120 incidents[12].

For future work, more data is needed to supplement the selected sample: the available data was scanty and not compatible for utilization in machine learning from the source Agency's website. This led to a lot of preprocessing in order for it to give the obtained results. Furthermore, there is need for addition of more variables during collection of the data since most of the variables obtained were general causations of fatalities and for better usable models to be achieved there is need to look into other factors for example, the terrain where the accidents occurred, road conditions and weather conditions.

6. REFERENCES

[1] Act Title: Traffic.2012. Kenya Law-Laws of Kenya. Retrieved from http://kenyalaw.org:8181/exist/kenyalex/actview.xql?actid=CAP.%20403.Accessed on 20th December,2019.

[2] Cai, B., Liur, Y., Liu, Z., Chang, Y and Jiang, R. 2020. Application of Bayesian Networks in Reliability Evaluation. Bayesian Networks for Reliability Engineering. Springer, Singapore.

[3] Cai, B., Huang, L. and Xie, M. 2017. Bayesian networks in fault diagnosis. IEEE Trans. Ind. Inf. 13(5), pp.2227–2240.

[4] Chakraborty, S., Mengersen, K. and Fidge, C. 2016. A Bayesian Network-based customer satisfaction model: a tool for management decisions in railway transport. Decis. Anal. 3, 4 doi:10.1186/s40165-016-0021-2.

[5] Chao, W., Xin, L., Zhunchen, H., Yakun and M., Wenjia.2019. Interpretable Charge Prediction for Criminal Cases with Dynamic Rationale Attention. Journal of Artificial Intelligence Research. 66. 743-764.10.1613/jair.1.11377.

[6] Daily Nation.2019. WHO: Kenya road deaths four times higher than NTSA reported. Retrieved from https://www.nation.co.ke/news/Kenya-road-deaths-

grossly-underreported–WHO/1056-4893792-ve7d07z/index.html. Accessed on 16th December,2019.

[7] Devin, S.2018. Introduction to Bayesian Networks. Retrieved from https://towardsdatascience.com/introduction-to-bayesian-networks-81031eeed94e. Accessed on 19th December,2019.

[8] Deublein, M., Schubert, M., Adey, T., Kohler, M. and Faber, H. 2013..Prediction of road accidents: A Bayesian hierarchical approach. Accident Analysis and Prevention.

[9] Microsoft. 2019. Normaliza Data. Microsoft. Retrieved fromhttps://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/normalize-data.Accessed on 19th December,2019.

[10] Mujalli, R., O., G., López and L., Garach. 2016. Bayes classifiers for imbalanced traffic accidents datasets. Accident Analysis and Prevention v. 88, p. 37-51.

[11] Nils, J., S. 2018. Artificial Intelligence-Bayes Network. Retrieved from https://www.norwegiancreations.com/

2018/09/artificial-intelligence-bayes-network/. Accessed on 19th December,2019.

[12] Richante.2014. What is the difference between a Bayesian network and a naïve Bayes classifier? Retrieved from https://stackoverflow.com/questions/12298150/what-is-the-difference-between-a-bayesian-network-and-a-naive-bayes-classifier.Accessed on 20th December,2019.

The Verification of Voice Recognition Using Cmusphinx and DTW

Gusti Made Arya Sasmita
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

I Putu Arya Dharmaadi
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

Henrico Aldy Ferdian
Department of Information Technology
Faculty of Engineering
Udayana University
Badung, Bali, Indonesia

Abstract: The development of smartphones increasingly affects the human life, such as data security. The use of smartphones to maintain data security is still low, which generally only uses conventional method as security systems available on smartphone devices, such as word combinations, PINs or patterns. Various weaknesses of conventional methods cause the development of biometrics systems. Therefore, a technology is created to replace the conventional security system that is biometric security system using natural human characteristics, one of them using voice. Data security uses the android based sound biometric app using CMUSphinx as the word library. CMUSphinx does not require an internet connection to run it. MFCC (Mel Frequency Cepstrum Coefficients) is as a characteristic extraction method on a digital sound signal. Matching process with saved data uses DTW (Dynamic Time Warping) method. This study can help improve the data security of smartphone; therefore, it cannot be sabotaged or accessed by other parties that are not desired.

Keywords: Voice Recognition, Android, CMUSphinx, MFCC, DTW

1. INTRODUCTION

The development of the internet is also followed by the development of an increasingly large smartphone, required the development of an adequate security system as well. Meanwhile, awareness on the data security of user is still low, which generally only use the conventional methods available on smartphones, such as word combinations, PIN or pattern. The use of PIN and password causes some problems, such as forgotten, can be used together, and can be cracked with various algorithms.

Various weaknesses of conventional methods above cause of the development of biometrics system. Users try to apply the new science of biometrics as a medium for personal recognition. Biometrics systems use body parts or behavior, which weaknesses in conventional methods can be reduced. The advantages of using biometrics are difficult to duplicate body parts, cannot be used together, and cannot be forgotten. This technology fulfills two important functions, they are identification and verification. The identification system aims to solve one's identity. Meanwhile the verification system aims to refuse or accept identity claimed by someone. The things that encourage the use of biometric identification and verification are universal (in everyone's case), unique (each has its own characteristics), and is not easily falsified[1].

Voice communication is one of the most rapid and appropriate communication media in humans in conveying information. The distinctiveness of the people' voice are in the loud or weak voice when people speak in normal circumstances, the way of pronunciation, intonation, rhythm of speech, accent etc. The use of sound is important to be analyzed in several processes related to voice processing which are divided into two types, namely speech recognition and speaker recognition. In contrast to speaker recognition which is the recognition of identity claimed by someone from his voice (special characteristic can be intonation of voice, depth of This research is applied in the Android platform because Android is one of the popular operating system used by the community. The used word library is CMUSphinx with the MFCC method as the feature extraction process on the voice level, etc.), speech recognition is a process done by computer to recognize words spoken by someone regardless of the related person identity[2].

This research is applied in the Android platform because Android is one of the popular operating system used by the community. The used word library is CMUSphinx with the MFCC method as the feature extraction process on the inputted digital sound pattern and the DTW method as the matching process. Library CMUSphinx is used because the security system designed does not require an internet connection to run it; therefore, users do not have to worry about the limitations of internet connection. The sound pattern feature matching process that has been stored in the database using the DTW method allows a device to recognize speech by digitizing words and matching those digital signals to a certain pattern stored in the device. This technology can be a good alternative for smartphones not easily sabotaged or accessed by other parties that are not desired.

The preparation of this paper is as follows: Section 2 presents some earlier work on speech recognition. Section 3 describes the proposed voice recognition system. The experimental results are discussed in Section 4 and conclusion in Section 5.

2. LITERATURE REVIEW

Research on voice recognition with several methods has been discussed several times. In 2011, Darma Putra and Adi Resmawan discuss how to design and create a software that can verify a speaker using MFCC method as feature extraction and DTW for matching process[3].

In 2013, B. Raghavendhar Reddy and E. Mahender discuss a system of acquiring speech signals that run through the microphone and processing sample speech to recognize spoken text. Recognized text can be stored in a file. The development is on the android platform using Eclipse Workbench. The speech-to-text system directly acquires and converts voice to text. It gives users different options for data entry. Furthermore, the speech-to-text system can improve the accessibility of the system by providing data entry options for blind, deaf, or physically disabled users[4].

In 2014, Bhadragiri Jagan Mohan and Ramesh Babu. N conducts research that explains the continuous speech recognition can be used in the security system to verify the keywords spoken by the user. The speech recognition system processes the spoken word using the MFCC algorithm and through feature matching stages using the DTW method. The whole system is implemented using matlab where the input of speech samples is recorded by sound card on windows[5].

In 2015, Mansour, et al. conducts a research on voice recognition using MFCC and DTW. This study focuses on developing a system for speech recognition using dynamic time warping (DTW) algorithms by comparing sound speaker signals with sound signals already stored in the database, and extracting the main features of speaker sound signals using MFCC[6].

3. RESEARCH METHOD

This study uses the Android platform with analog voice signal processing using CMUSphinx as the word library. The analog voice signal is converted to a digital voice signal; furthermore it is processed using the MFCC method as feature extraction. Sample data already registered, will be tested for data matching using DTW method. This allows the system to know which users match the data that has been registered or not. Figure 1 presents the block diagram of the created application.

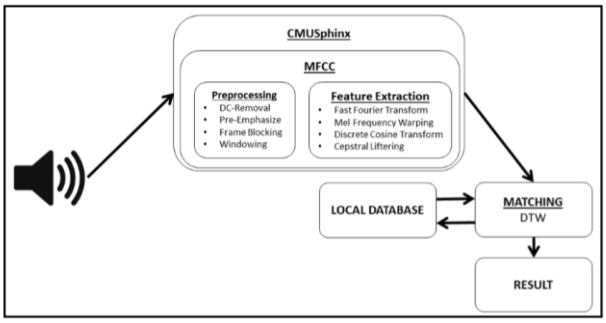


Figure 1: Block Diagram

4. CONCEPTS AND THEORIES

Concepts and theories contain explanations of supporting theories that will be used in this study. These theories include Voice Recognition, CMUSphinx, MFCC, DTW. The theory will be discussed as follows.

4.1 Voice Recognition

Sound is a combination of various signals; however, the pure sound can theoretically be explained by the velocity of the oscillation or frequency measured in Hertz (Hz) and the amplitude or loudness of the sound by measurement in decibels (dB). Voice recognition first appeared in 1952 and consists of a device for the introduction of a single digit of spoken word. Then in 1964, there is IBM Shoebox, one of the most well-known technology in the United States in the field of health is Medical Transcriptionist (MT) is a commercial application that uses speech recognition. Voice recognition is divided into two types, namely speech recognition and speaker recognition. Speech recognition is the process of voice identification based on the spoken word. The comparable parameter is the level of voice suppression which will then be matched to the available database templates. Meanwhile the voice recognition system based on people who speak called speaker recognition. User recognition can be classified into three stages: identification, detection, and verification. User identification is a process for determining the identity of a user through spoken voice, meanwhile user detection is the process of discovery of the user's voice from a bunch of votes in a database, and user verification is a process to verify the user's voice conformance to the identity claimed by the user. User recognition focuses more on user voice recognition and not on user speech recognition[7].

4.2 CMUSphinx

In the early 1920s the first machine to recognize a commercially significant level of speech called Radio Rex was personalized in the 1920s, the effort to design speech recognition systems was automatically made in the 1950s. During the 1950s, most speech recognition systems investigated the resonance spectral vowel regions of each speech extracted from the analog filter output signal and the logic circuit[8].

In 2000, the Sphinx group at Carnegie Mellon is committed to opening some voice recognition components, including Sphinx 2 and later. Sphinx 3, Sphinx 4, PocketSphinx for mobile devices. PocketSphinx as an open source library allows developers to add new languages. However, it takes an acoustic model and a language model. The use of PocketSphinx technology in 2014 by P.Vijai Bhaskar and Dr. S. Rama Mohana Rao came with the Telugue welcome recognition system and there was little effort made in translating the Tamil language into English by voice. Sphinx is an opensourcetoolkit for speech recognition developed by Carneige Mellon University (CMU) located in the United States. In order to recognize and respect the creator, the Sphinx is often referred to as CMUSphinx. CMUSphinx uses the HMM method and the n-gram statistical language model to build an Automatic Speech Recognition (ASR) system. CMUSphinx was first developed by Kai-Fu Lee[9].

4.3 Mel Frequency Cepstrum Coefficients (MFCC)

Mel Frequency Cepstrum Coefficients (MFCC) is one of the most widely used methods in the field of speech processing, both speech recognition and speaker recognition used to perform feature extraction. This method adopts the workings of the human auditory organ, so as to capture the very important sound characteristics which are used to perform parameter extraction, a process that converts voice signals into several parameters. The parameter extraction steps using the MFCC method is as shown in Figure 2[3].

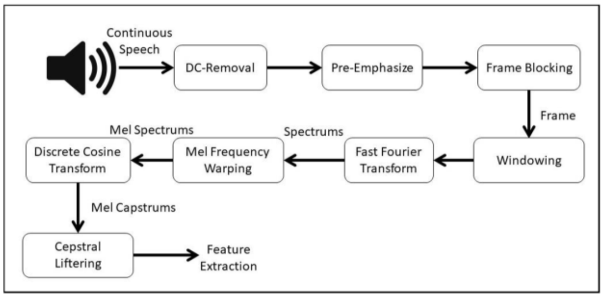


Figure 2: Block MFCC

4.3.1 Conversion of Analog Signals Into Digital Signals

Natural signals in general such as voice signal is a continuous signal which has an unlimited value. On the computer, all the signals that can be processed by the computer is just a discrete signal or often known as the term digital signal. Changing the signal can be done through 3 processes, such as data sampling process, quantization process and coding process[3].

4.3.2 DC-Removal

DC Removal aims to calculate the mean of the sound sample data, and subtract the value of each sound sample by the average value. The goal is to get the normalization of the input voice data.

$$y[n] = x[n] \cdot \bar{X} \ , 0 \leq n \leq N\text{-}1 \qquad (1)$$

Where:
y[n] : The signal sample of DC removal process results.
x[n] : The original signal sample.
$\bar{X}$: The average value of the original signal sample.
N : The signal length.

4.3.3 Pre-Emphasize

Pre-emphasize is one type of filter that maintains high frequencies on a spectrum, which is generally eliminated during the sound production process. The x (n) speech signal is sent to the high-pass filter[10]:

$$y(n) = x\ (n) - a * x\ (n - 1) \qquad (2)$$

Where y (n) is the output signal and the value of a is usually between 0.9 and 1.0. The Z transform of this equation is given by:

$$H(z) = 1 - a * z^{-1} \qquad (3)$$

The purpose of pre-emphasize is to offset the high frequency parts suppressed during the human voice production mechanism. Furthermore, it can reduce the noise ratio of the signal in order to improve signal quality.

4.3.4 Frame Blocking

Frame blocking is a process in which the voice signal is divided into several pieces that can later facilitate the calculation and analysis of sound. Each piece of the sound signal is called a frame. Frames in frame blocking generally have a length of 10-30ms. One frame consists of several samples depending on each second the sound will be sampled, how big the sampling frequency and overlapping. Overlapping is done to avoid loss of feature or voice characteristics on the boundary of intersection of each frame. The length of the overlap area commonly used is approximately 30% -50% of the frame length.

4.3.5 Windowing

Framing process can cause spectral or aliasing leakage. This can happen because of the low number of sample rate, or due to the frame blocking process which causes the signal to be discontinue. In order to reduce the possibility of spectral leakage, the result of the farming process must pass through the windowing process. Each frame must be multiplied by a hamming window to keep the first and last point continuity in the frame called Hamming Window.

4.3.6 Fast Fourier Transform (FFT)

The core of the fourier transform is to decipher the signal into the component of the sine shape of different frequencies. Spectral analysis shows that the time difference in speech signal is related to different energy distributions over frequency. Therefore, FFT is done to get the frequency response of each frame. When FFT is done on a frame, it is assumed that the signal in the frame is periodic and continuous when it is enclosed. If this does not happen, FFT can still be performed; however, discontinuities on the first frame and last point will likely introduce undesirable effects on the frequency response. In order to solve this problem, the user multiplies each frame by windowing to increase its continuity on the first and last points. FFT is a fast algorithm for the implementation of Discrete Fourier Transform (DFT) operated on a discrete time signal consisting of N sample as follows:

$$f(n)=\sum_{K=0}^{N-1} y_k e^{-2\pi jkn/N}, n=0,1,2,\ldots,N\text{-}1 \qquad (4)$$

4.3.7 Mel Frequency Warping

Psychophysical studies have shown that the human perception of sound frequencies for speech signals does not follow a linear scale. Therefore, for any tone with real frequency f, in Hz, a pattern is measured on a scale called 'mel'. The 'mel frequency' scale is a linear frequency scale below 1000 Hz and a logarithmic scale above 1000 Hz. This scale is defined by Stanley Smith, John Volkman and Edwin Newman as[11]:

$$mel(f)=2595*\log_{10}(1+\frac{f}{700}) \qquad (5)$$

An approach for spectrum simulation on a mel scale is to use a filterbank placed uniformly on a mel scale. Filterbank is one of the forms of the filter performed in order to know the energy size of a particular band frequency in the sound signal. At MFCC, filterbank is applied in the frequency domain.

4.3.8 Discrete Cosine Transform (DCT)

DCT is the last step of the main process MFCC feature extraction. Basically the DCT concept is the same as the inverse fourier transform. However, the results of DCT close to PCA (Principle Component Analysis). PCA is a classic static method that is widely used in data analysis and compression. Therefore, DCT often replaces the inverse fourier transform in the MFCC feature extraction process. In order to get the value of MFCC cepstrum, the mel frequency must be transformed back into time domain using Discrete Cosien Transform (DCT) method. DCT is applied to the output of a triangular N bandpass filter in order to obtain a coefficient of cepstral L mel-scale. The formula for DCT is,

$$C(n) = \Sigma\ Ek * \cos(n * (k - 0.5) * \pi / 40)) \qquad (6)$$

Where n = 0,1, .. to N

Where N is the number of triangle bandpass filter, L is the number of cepstral mel-scale coefficients. There are N = 40 and L = 13. Since it has done FFT, DCT converts the frequency domain into a domain like time called domain quefrency. The features obtained are similar to cepstrum, which is called the coefficient of cepstral mel-scale, or MFCC.

4.3.9 Cepstral Liftering

The results of the DCT process have some disadvantages. Low order from cepstral coefficients is very sensitive to spectral slope; meanwhile the high order part is very sensitive to noise. Therefore, cepstral liftering becomes one of the standard techniques applied to minimize the sensitivity. This process can be done by implementing the window function to the cepstral features.

$$W[n] - \{1+\frac{L}{2}\sin(\frac{n\pi}{L})\ n - 1,2,...,L \quad (7)$$

Where:
L = number of cepstral coefficients.
N = index of cepstral coefficients.

4.4 Dynamic Time Warping (DTW)

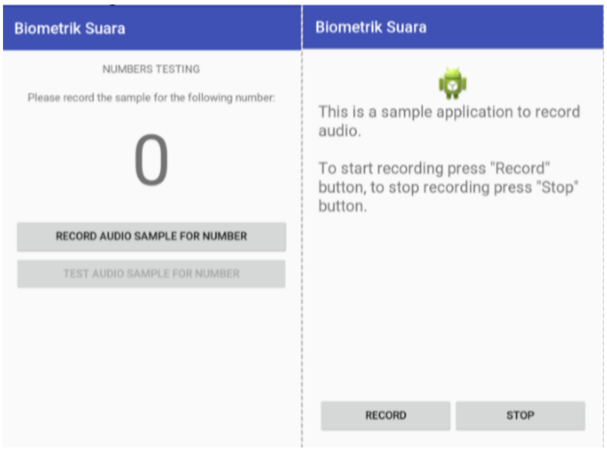
The problems in voice recognition are quite numerous; one of them is the recording process that is often different in duration, even if the spoken word or phrase is the same. Although for the same syllable or vowel, the recording process often occurs in different durations. Consequently the matching process between test signals and reference signals (templates) often does not yield optimal values. DTW (Dynamic Time Warping) is an algorithm that focuses on matching two feature vector sequences by repeatedly shrinking or extending the time axis until an exact match is obtained between two sets. It is used in order to check the similarity between two voice signals or non-linear curved time series. The DTW distance between two vectors is calculated from the optimal bending path of the two vectors.

The DTW algorithm is more realistic to use in measuring pattern matching than simply using linear measurement algorithms such as Euclidean Distance, Manhattan, Canberra, Mexican Hat and others[12]. The principle provides a range of 'steps' in space (a time frame in the sample, time frames in the template) and is used to match paths that show the largest local match (similarity) between straight time frames. The total similarity cost obtained with this algorithm is an indication of how well these samples and templates have in common which will be selected as the bestmatching templates.

5. EXPERIMENT AND RESULT

5.1 Experiment

Figure 3 is an app view that has been created for testing. The sample used is 50 different person voice sample data by performing tests on three conditions around a silent environment, noise with low level, and noise with high level.
The application created will be tested by looking for an error ratio that states the probability of a matching error in the system. There are two types of ratios, namely the False Accepted Rate (FAR) and False Rejected Rate (FRR) ratios. The testing phase in this study was carried out by a total of 50 people who were asked to pronounce a word from a given word list where everyone has one word to register and test. Furthermore, the user is asked to say the words given to the conditions surrounding the environment of registration and testing sounds silent, noise with low level and high level and matching results recorded.



Biometrik Suara

NUMBERS TESTING

Please record the sample for the following number:

0

Biometrik Suara

This is a sample application to record audio.

To start recording press "Record" button, to stop recording press "Stop" button.

Figure 3: Application Interface

5.2 Result

5.2.1 Low Noise Condition Testing

Low noise condition testing is the result of matching when the environment of noise conditions is heard with low levels, such as the sound of small children's cries from a distance, people chatting from a distance, etc. At registration and testing as well as recording results, out of 50 users are divided into 5 groups in which each group has 10 members. In speech recognition, the test is repeated 3 times with the same word of each user and the matching results are recorded.

Test results on conditions around the test environment have low audible noise, in each group resulting in average variation of recognition as well as FRR values and FAR values obtained.

Group 1 gets an average introduction of 100% with an FRR value and a FAR value of 0%. Group 2 gets an average introduction of 60% with a FRR value of 40% and a FAR value of 20%. Group 3 gets an average introduction of 80% with a FRR value of 20% and a FAR value of 10%.

Group 4 gets an average introduction of 90% with FRR value and a FAR value of 10%. Meanwhile group 5 gets an average introduction of 60% with a FRR value of 40% and a FAR value of 30%. The graph of accuracy of each group on the environment condition sounds low-level noise is shown in Figure 4.

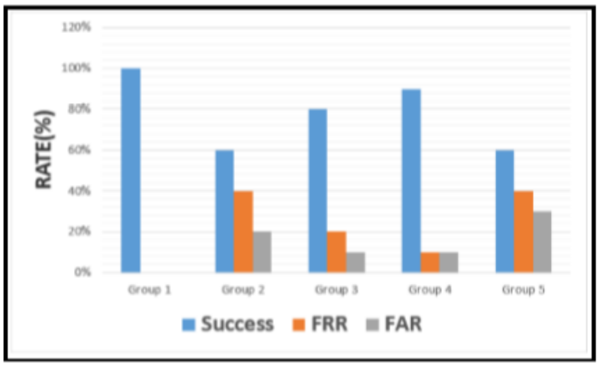


Figure 4: The Graph of Accuracy in Each Group

Figure 4 is a graph of accuracy in each group, in which the success rate of the system in the test can be said to be successful in voice recognition because the average introduction in each group can be said evenly, that is 60% to 100% which there is no result too low

5.2.2 High Noise Condition Testing

High noise condition testing is a test where conditions around the test is noisy and noise is caused as a child is playing closely, neighbors are turning on the sound system too loud, etc. The first test, under high environmental noise condition, is carried out on 50 users with one test. The second test, from 10 users taken in sequence to sample data entered in the database cleaned and removed from the noise and tested using noise obtained previously. Noise removal and cleaning is done using the Adobe Audition CS6 application. The following test scores for each user are shown in Table 1.

Table 1: Test Score 1 High Noise Condition.

No.	Kata	Klasifikasi Benar	Klasifikasi Salah	Rata-rata Pengenalan
1	Satu	1	0	100%
2	Dua	0	1	0%
3	Tiga	1	0	100%
4	Empat	0	1	0%
5	Lima	1	0	100%
6	Enam	1	0	100%
7	Tujuh	0	1	0%
8	Delapan	0	1	0%
9	Sembilan	1	0	100%
10	Sepuluh	0	1	0%
11	Sebelas	1	0	100%
12	Dua Belas	0	1	0%
13	Tiga Belas	0	1	0%
14	Gajah	0	1	0%
15	Lima Belas	0	1	0%
16	Enam Belas	0	1	0%
17	Tujuh Belas	0	1	0%
18	Delapan Belas	1	0	100%
19	Sembilan Belas	0	1	0%
20	Dua Puluh	0	1	0%
21	Dua Puluh Satu	0	1	0%
22	Dua Puluh Dua	1	0	100%

23	Dua Puluh Tiga	1	0	100%
24	Dua Puluh Empat	1	0	100%
25	Dua Puluh Lima	0	1	0%
26	Dua Puluh Enam	1	0	100%
27	Lima Puluh Enam	1	0	100%
28	Tujuh Puluh Delapan	0	1	0%
29	Dua Puluh Sembilan	1	0	100%
30	Tiga Puluh	1	0	100%
31	Domba	1	0	100%
32	Tiga Puluh Dua	1	0	100%
33	Tiga Puluh Tiga	0	1	0%
34	Tiga Puluh Empat	0	1	0%
35	Tiga Puluh Lima	0	1	0%
36	Tiga Puluh Enam	1	0	100%
37	Sembilan Puluh	0	1	0%
38	Dua Ratus Lima Puluh	1	0	100%
39	Seratus	1	0	100%
40	Delapan Puluh Tujuh	0	1	0%
41	Empat Puluh Satu	0	1	0%
42	Enam Puluh Tujuh	1	0	100%
43	Lima Puluh Lima	1	0	100%
44	Empat Puluh Empat	1	0	100%
45	Empat Puluh Lima	1	0	100%
46	Delapan Puluh Sembilan	0	1	0%
47	Enam Puluh Tiga	1	0	100%
48	Enam Puluh Sembilan	0	1	0%
49	Empat Puluh Sembilan	1	0	100%
50	Lima Puluh	1	0	100%
TOTAL		**26**	**24**	**52%**
Nilai FRR				**48%**
Nilai FAR				**32%**

Table 1 is the first test of high noise conditions where 50 user data are tested once test and high noise results result in less accurate matching with an average acquired 52%. The result is a FRR value of 48%, meanwhile for FAR value of 32%.

Table 2: Test Score 2 High Noise Condition.

No.	Kata	Klasifikasi Benar	Klasifikasi Salah	Rata-rata Pengenalan
1	Satu	3	0	100%
2	Dua	3	0	100%
3	Tiga	-	-	-
4	Empat	0	3	0%
5	Lima	0	3	0%
6	Enam	0	3	0%
7	Tujuh	3	0	100%
8	Delapan	3	0	100%
9	Sembilan	-	-	-
10	Sepuluh	3	0	100%
TOTAL		**15**	**9**	**62,5%**
Nilai FRR				**30%**
Nilai FAR				**30%**

Table 2 is a test of 10 sample data where 10 sampled data entered in the database are cleared and removed from noise and tested using data containing noise. The average introduction is 62.5% with FRR value and 30% FAR value. The result is not good because the effect of noise, user 3 and user 9 looks empty because when tested with high noise, the application does not find the matching data in the database entered. However, when the test data is done noise reduction of 100% using Adobe Audition CS6 application and re-tested; therefore, the data is not detected initially or wrong will rematch with existing data in the database.

5.2.3 The Silent Condition Testing

Testing data is carried out when conditions is quiet, the absence of noise interference from near or far, such as sound system, screams small children etc. Test is performed on 30 users by means of every 10 users who have registered and tested, then reenrolled 10 users until 30 users are enrolled and tested. The test graphs in silent environment conditions are described in Figure 5.

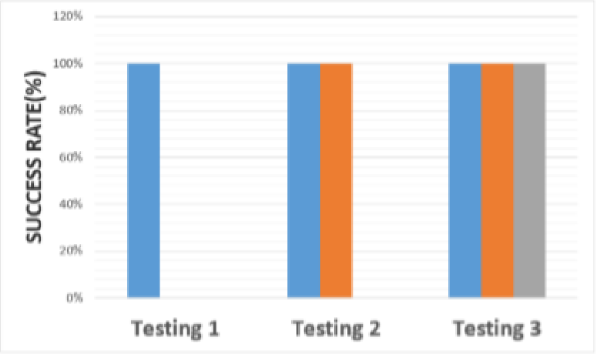


Figure 5: Graph of Accuracy Gradual Testing on Silent Condition

Figure 5 is a gradual testing graph in a quiet state. The results obtained are said to be successful because of all the tests at the time of silence the average introduction of 100% with the value of FRR and FAR value of 0%. 5.2.4 The accuracy comparison results of any environmental condition The accuracy comparison results of any environmental condition at registration or test are presented in the form of an accuracy comparison graph described in Figure 6.

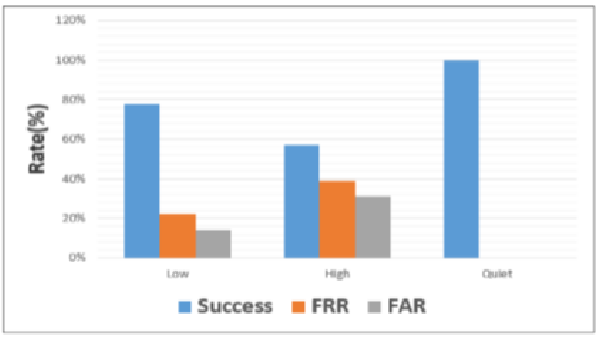


Figure 6: Graph of Comparison of Accuracy of Each Condition

Figure 6 is the average accuracy of the system during environmental conditions around the test with low noise conditions with a low level of 78%, 57.25% of high audible noise conditions and a 100% silent sound condition. Through the comparison graph on each condition, it can be concluded that noise on the environment is one of the factors that have an important effect on the quality of the system in the results of matching test data. Other factors can affect the system in the introduction of testing:

a. Environmental conditions.
b. Conditions and intonation of the user's voice.
c. Location of Microphone.
d. Ways of recording a voice signal.
e. Equipment condition.

6. CONCLUSION

Voice recognition is one type of the biometrics introduction. The use of the CMUSphinx library is as a part of an Androidbased Biometrics Application system that runs without an internet connection and it has a good performance in converting the analog voice signal into a digital sound signal using the MFCC method. Matching using the DTW method is a flexible mathematical method, giving high accuracy results. The best result is obtained when condition is quiet around the test environment, with all the tests matching the data stored in the database and the FRR value and the FAR value of 0%. Testing at a time when noise condition is heard with low levels can be said to be successful because almost all the tests detected match the data stored. Meanwhile, the bad result is when testing at highaudible noise conditions with a FRR value of 48% and a FAR value of 32%.

REFRENCES:

[1] K. Agustini, “Biometrik Suara dengan Transformasi Wavelet Berbasis Orthogonal Daubenchies,” Gematek J. Tek. Komput., vol. 9, no. 1, pp. 49–57, 2007.

[2] G. Melissa, “Pencocokan Pola Suara (Speech Recognition) dengan Algoritma FFT dan Divide and Conquer,” Makal. If2251 Strateg. Algoritm., 2008.

[3] D. Putra and A. Resmawan, “Verifikasi Biometrika Suara Menggunakan Metode MFCC Dan DTW,” Lontar Komputer, vol. 2, no. 1, pp. 8–21, 2011.

[4] B. R. Reddy and E. Mahender, “Speech to Text Conversion using Android Platform,” Int. J. Eng. Res. Appl., vol. 3, no. 1, pp. 253–258, 2013.

[5] B. J. Mohan and N. Ramesh Babu, “Speech recognition using MFCC and DTW,” 2014.

[6] A. H. Mansour, G. Z. A. Salh, and K. A. Mohammed, “Voice Recognition using Dynamic Time Warping and MelFrequency Cepstral Coefficients Algorithms,” Int. J. Comput. Appl., vol. 116, no. 2, pp. 34–41, 2015.

[7] C. Ho, Speaker Recognition System. California: California Institut of Technology, 1998.

[8] L. R, A. S, S. S, B. C, and G. Fernando, “Android Speech-to-speech Translation System for Sinhala,” Int. J. Sci. Eng. Res., vol. 6, no. 10, pp. 1660–1664, 2015.

[9] I. K. Suryadharma, G. Budiman, and B. Irawan, “Perancangan Aplikasi Speech To Text Bahasa Inggris ke Bahasa Bali Menggunakan Pocketsphinx Berbasis Android (Design Application Speech To Text English To Balinese Language Using Pocketsphinx Base On Android),” Bandung Univ. TELKOM, pp. 1–10, 2014.

[10] K. Chakraborty, A. Talele, and P. S. Upadhya, “Voice Recognition Using MFCC Algorithm,” Int. J. Innov. Res. Adv. Eng., vol. 1, no. 10, pp. 2349–2163, 2014.

[11] A. Setiawan, A. Hidayatno, and R. R. Isnanto, “Aplikasi Pengenalan Ucapan dengan Ekstraksi Mel-Frequency Cepstrum Coefficients (MFCC) Melalui Jaringan Syaraf Tiruan (JST) Learning Vector Quantization (LVQ) untuk Mengoperasikan Kursor Komputer,” TRANSMISI, vol. 13, no. 3, pp. 82–86, 2011.

[12] A. Muhammad, “Penggunaan Jarak Dynamic Time Warping (DTW) pada Analisis Cluster Data Deret Waktu (Studi Kasus pada Dana Pihak Ketiga Provinsi Seindonesia),” pp. 277–280, 2005.