Feature selection for DDoS detection using classification machine learning techniques

Andi Maslan[1], Kamaruddin Malik Bin Mohamad[2] Feresa Binti Mohd Foozy[3]
[1]Teknik Informatika, Universitas Putera Batam, Jl. Letjen Soeprapto, Muka Kuning Batam, 29432, Indonesia
[2,3]University Tun Hussein Onn Malaysia UTHM, Parit Raja, Johor Baru, 86400, Malaysia

Article Info

Article history:

Received Nov 17, 2019
Revised Feb 12, 2020
Accepted Feb 19, 2020

Keywords:

Classification
DDoS
Feature Selection
Machine Learning
Network Security

ABSTRACT

Computer system security is a factor that needs to be considered in the era of industrial revolution 4.0, namely by preventing various threats to the system, as well as being able to detect and repair any damage that occurs to the computer system. DDoS attacks are a threat to the company at this time because this attack is carried out by making very large requests for a site or website server so that the system becomes stuck and cannot function at all. DDoS attacks in Indonesia and developed countries always increase every year to 6% from only 3%. To minimize the attack, we conducted a study using Machine Learning techniques. The dataset is obtained from the results of DDoS attacks that have been collected by the researchers. From the datasets, there is a training and testing of data using five techniques classification: Neural Network, Naïve Bayes and Random Forest, KNN, and Support Vector Machine (SVM), datasets processed have different percentages, with the aim of facilitating in classifying. From this study it can be concluded that from the five classification techniques used, the Forest random classification technique achieved the highest level of accuracy (98.70%) with a Weighted Avg 98.4%. This means that the technique can detect DDoS attacks accurately on the application that will be developed.

Corresponding Author:

Andi Maslan,
Department of Informatic,
Putera Batam University,
Jl Letjen Soeprapto, Muka Kuning Batam, Kepulauan Riau, Indoneisa.
Email: Lanmasco@gmail.com

1. INTRODUCTION

The computer security system is a factor that needs to be considered in the era of industrial revolution 4.0, namely by preventing various threats to the system, as well as being able to detect and repair due to any damage that occurs. According to [1], broadly the threat to information systems can be divided into two types, namely active threats and passive threats. Active threats include fraud and crimes against computers, while passive threats include system failure, human error, and natural disasters. System failure states failure in component equipment such as hard disk or computer network itself. From this concept, computer-based systems and networks sometimes become vulnerable to fraud and data theft. One type of attack that still exists and difficult to stop is the Distributed Denial of Services (DDoS) attack. This attack is carried out by making many requests for a site or website server so that the system becomes stuck and cannot function at all. Another attack that is also very dangerous is a sniffer attack technique. This technique is implemented by creating a program that tracks someone's data packet when the packet crosses the internet, captures passwords or captures contents. And that is not less important is the technique of spoofing attacks, this technique is done by falsifying e-mail addresses or the web in order to trap users to enter important information such as passwords or credit card numbers. Of the various types of learning, this study focuses on DDoS attacks.

DDoS attacks in Indonesia have been increasing recently. Data showed that 79% of total DDOS attacks in the fourth quarter of 2017 were intended for game applications. Actually, the figure fell three percent compared to the attack in the previous quarter. While telecommunication and internet applications increased to 6% from only 3%, also the application of financial services rose 2% to 4% in the last quarter of 2017. DoS is a form of DoS attack when an attacker makes the network inaccessible (slowing down or losing data) by attacking using more than one Protocol (IP) Internet address. This causes a flood of traffic making it difficult to identify the attacker. DDoS attacks are very detrimental both operationally and financially. In the B2B International survey in collaboration with Kaspersky Lab, entitled Global Corporate IT Security Risks 2015, it can identify that a DDoS attack on an online resource can cause financial losses starting at the US $ 53-417 thousand.

To anticipate attacks by network security, researchers always looking for the best techniques for detecting DDoS attacks, such as research conducted [2], how to detect DDoS attacks by developing statistical-based DDoS detection systems using Multivariate Correlative Analysis (MCA). MCA uses the Triangle-Area-Map (TAM) representation technique to describe the relationship between each traffic feature by calculating the distance of one feature value to another feature value for each feature extracted. Data from MCA processing results were analyzed by using Mahalanobis Distance to be used as reference or observation data. The detection process of the observed threshold-based data from the reference data and the anomaly classification process using Mahalanobis Distance and Cosine Distance to calculate the distance between the values of the TAM traffic feature observed with the TAM reference traffic. System testing was done by measuring the accuracy of the algorithm, based on the results of the system with parameters Detection Rate (DR), False Positive Rate (FPR) and Accuracy (ACC).

In research [3], in his research developed a detection method by looking at DDoS attack patterns using network packet analysis and utilizing machine learning techniques to study DDoS attack patterns. In his research, to analyze a large number of network packages provided by the Applied Internet Data Analysis Centre and implement a detection system using Vector Machine Support (SVM) with radial (Gaussian) kernel basic functions. Accurate detection system for detecting DDoS attacks. While the results of the study [4] explained that the attackers (hackers) can do more DOS attacks with zombie hosts (computers that have been injected with the remote control script/botnet) on targets distributed and simultaneously so that the effect of this attack is an ability to knock out the target quickly. Based on a number of studies, the CUSUM algorithm is recognized as having an accuracy point that is quite reliable in detecting DDOS attacks that often occur today. UDP Flood attacks also dominated several major attacks in the world. Based on the problem of the fact that the UDP flood dominates the current attacks, the author wanted to create an IDS (Intrusion Detection System) using the CUSUM algorithm. It is expected that the application of the CUSUM algorithm on the IDS system is able to detect UDP Flood attacks by approaching high accuracy and fast detection time. In research [5] aimed to develop a new approach to detect DDoS attacks, based on network logs that were statistically analyzed with the function of the neural network as a detection method. Training data and testing were taken from CAIDA DDoS Attack 2007 and independent simulations. Testing of statistical analysis methods on network logs with neural network functions as detection methods resulted in an average percentage of recognition of three network conditions (normal, slow DDoS, and DDoS) of 90.52%. The new approach to detect DDoS attacks was expected to be a complement to the Intrusion Detection System (IDS) system in predicting DDoS attacks.

In research [6-7] the byte level analysis of HTTP traffic offers a practical solution to the problem of network intrusion detection and traffic analysis. Such an approach does not require any knowledge of applications running on web servers or any pre-processing of incoming data. In this project, he applied three N-gram based techniques to the problem of HTTP attack detection. The goal of such techniques was to provide the first line of defense by filtering out the vast majority of benign HTTP traffic. This technique in terms of accuracy of attack detection and performance. Techniques provide more accurate detecting and are more efficient in comparison to a previously analyzed HMM-based technique.

Research conducted by [3] developed an intelligent system for detecting DDoS attack patterns using network packet analysis and utilizing machine learning techniques to study DDoS attack patterns. In this study, Klyuev analyzed a large number of network packages provided by the Applied Internet Data Analysis Centre and implemented a detection system using SVM with a radial Kernel (Gaussian) base function. This research prepared three types of datasets that Klyuev used with three and five features. Detection system was more than 85% accurate with all types of datasets and 98.7% accurate with five features. The strategy for developing DDoS attack detection systems showed that system detection with SVM was trained using the proposed feature to successfully detect DDoS attacks with high accuracy.

In [8] that Fast Entropy and flow-based showed a significant reduction in computational time compared to conventional Entropy computation while maintaining good detection accuracy. The network traffic was analyzed and fast the entropy of requests per-flow was calculated. The DDoS attack was detected when the difference between the entropy of flow counts and the mean value of entropy in that time interval was that the threshold value was updated adaptively based on traffic pattern conditions to improve the detection accuracy. In detecting DDoS attacks this research proposed three methods, namely fast Entropy, flow aggregation, and adaptive Threshold.

In [9] this paper, he collected a new dataset that included modern types of attacks, which were not used in previous research. The dataset contained 27 features and five classes. A network simulator (NS2) was used in this work because NS2 could be used with high reliability and reasonable results that reflected a real environment. In [10-12] Attack or intrusion into a system is something that is almost certainly happened in the world nowday of information technology. To overcome this, there are several technologies that can be used, such as firewalls or intrusion detection systems (IDS). Unlike firewalls that only inspect incoming packets based on IP address and port, IDS work by monitoring the payloads of the packet that come into a computer to then decide whether the incoming packet is malicious or not. An example of IDS application is Snort IDS, an open-source application that uses string matching to detect malicious activity. One weakness of string-matching IDS is the occurrence of a string in a packet must be an exact match, just a slight difference can make an attack comes undetected, making it difficult to detect attacks that have similar flow but different pattern. Therefore, this paper proposed an intrusion detection method using n-gram and cosine similarity to seek similarity of a couple of packet sequences, thus the searching is conducted by looking for the similarity between payload and existing signature. In contrast to Snort, those packets are not matched with the pattern of attacks, but rather the pattern of legitimate access to a web page done by legitimate users, so packets that have a high similarity are regarded as benign, while the low ones will be regarded as an attack. From the test results with a different value of the threshold, then we obtained the value of 0.8 with n = 3 gave the best accuracy. This intrusion detection system is also capable of detecting various types of attacks without having to define existing attacks in advance, making it more resistant to zero-day attacks.

According to the research conducted by [5], [13-14] that Distributed denial-of-service (DDoS) is an attack-type in which volume, intensity, and mitigation costs continue to rise with a growing scale of organization. This study has the objective to develop a new approach to detect DDO attacks, based on the characteristics of network activity using a neural network with the functionality of fixed moving average windows (FMAW) as a detection method. Data were taken from the training and testing of DDoS Attack Caida 2007 and standalone simulation. Testing of methods produced the detection percentage of three network conditions (normal, slow DDoS, and DDoS) amounted to 90.52%. A new approach in detecting DDS attacks, a system that predicts the occurrence of DDS attacks.

In [15]. This study classifies network traffic information which contains botnets using the K-Nearest Neighbour algorithm. The algorithm calculates the distance on each feature in the dataset and then identifies the type of flow based on the majority of certain neighbor values (k values). The test results in this study are 92.57% where the k value is determined according to the system default, namely 5. The best k value in this study cannot be determined because the test is done to determine the value of k to get a result with a difference in value that is quite far.

From the problems that have been described, the problem to be solved in this study is to address the number of features in the dataset so that it can find out the number of features that are most important in detecting DDoS attacks. To find out the level of detection of DDoS attacks, this study uses Classification Machine Learning Algorithms such as Naive Bayes, neural networks, SVN, KNN, and Random Forest. Of the five algorithms used, the expected end result is to be able to compare which algorithm is most accurate in detecting DDoS attacks with the features selected.

2. RESEARCH METHOD

In this study the dataset used was data obtained from research [9], the dataset in the study was 734,627, while in this study the dataset used for training was 5899 and for testing as many as 1770. The steps in this study were as follows and Figure 1 shows research process:

- Data collection is carried out in an on-going network that is captured using Wireshark.
- The data is then converted to CSV
- Feature Selection model regression
- Attributes that are not used will be fixed; attributes that are not used will be removed.
- After that, an analysis using a data mining tool will be analyzed and use some algorithm machine learning

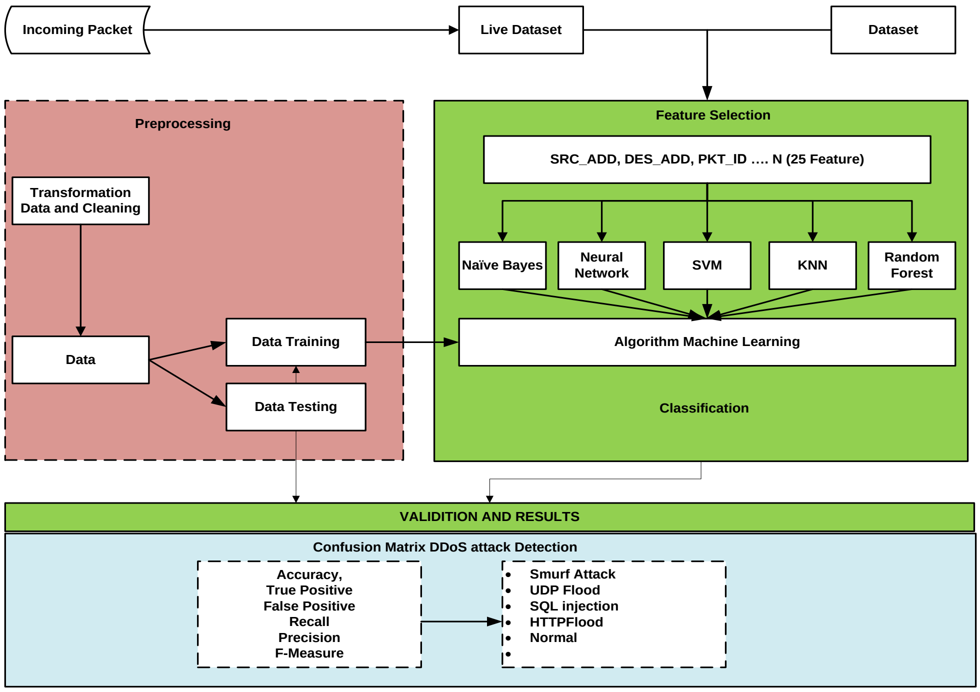

Figure 1. Research process

From this step, it can be seen that the classification technique used to detect DDoS attacks (Smurt Attack, UDP Flood, SQL Injection and HTTP Flood) and Normal Packets uses 5 classification models namely Naïve Bayes, Neural Network, SVM, KNN, and Random Forest. To find out the accuracy of detection, the parameters TF, FP, Recall, Precision, and F-Measure are used.

2.1. Dataset

The features used in this study were 25 features obtained from the results of a real-time attack simulation carried out for 3 days with 4 hours each visit. When simulating the number of feature attacks as many as 27 features [16], then extracting them using the Canadian Institute for Cybersecurity CICFlowMeter-V3 online application, because there are too many features, features that have the appropriate value are expected to be removed. The method used to maximize the features using the Regression model with SPSS applications, thus the features used for the training and testing process in this study are as follows on Table 1.

Table 1. Feature of dataset

Variable No	Feature Selection	Type	Variable No	Feature Selection	Type
1	Source Address	String	13	Packet in	String
2	Destination Address	String	14	Packet out	String
3	Packet ID	String	15	Packet Transmition	String
4	From Node	String	16	Packet delay note	String
5	To Node	String	17	Packet Rate	String
6	Packet Type	String	18	byte rate	String
7	Packet Size	String	19	Pkt Avg Size	String
8	Squencial Number	String	20	Utilization	String
9	Number of Packets	String	21	Packet Delay	String
10	Number of bytes	String	22	Packet send time	String
11	Node name from	Symbolic	23	Packet reserved time	String
12	Node Name To	Symbolic	24	The first packet Sent	String
			25	the last packet reserved	String

2.2. Feature selection

To find out the most optimal feature value in detecting DDoS attacks, dataset analysis is used to use linear regression with the forward method. In terms of mutual information, the purpose of feature selection is to find a feature set S with m features {xi}, which jointly has the largest dependency on the target class c. This scheme, called Max-Dependency, has the following formula [17].

$$\max D(S,c),\quad D = I(\{x_i, i = 1,\dots,m\}; c). \tag{1}$$

Obviously, when m equals 1, the solution is the feature that maximizes I {xj;c} (1<= j <= M}. When m > 1, a simple incremental search scheme is to add one feature at one time: given the set with m-1 features, $S_{m\text{-}1}$, the mth feature can be determined as the one that contributes to the largest increase of I{S;c}.

2.3. Algorithms machine learning

2.3.1. Naïve bayes

Naive Bayes Classifier is a collection with a statistical model for calculating classes that have each group of attributes that exist, and determine which class is the most optimal. In this method, all attributes will contribute to decision making, with the same important importance weights and each attribute is independent of each other [18]. The equation of the Bayes theory is:

$$P(H|X) = \frac{P(X|H.P(H)}{P(X)} \tag{2}$$

X: Data with classes that haven't known
H: Data hypothesis is a specific class
P(H|X): The probability of hypothesis H is based on condition X (prior probability)
P (H): Probability of hypothesis H (prior probability)
P (X|H): Probability X based on condition on the hypothesis H
P (X): Probability X

2.3.2. Random forest

Random forest is an ensemble learning method that was first proposed by [19] which is a combination of classification trees in such a way that each tree depends on the random value of the sample vector independently and with the same distribution for all trees in the forest. Random Forest has been widely used both for classification and regression because of its superior performance and simple structure. To handle unbalanced data, the RF algorithm undergoes a slight modification in the selection of training data, namely by balancing the number of records in the major and minor classes. This technique is called Balanced Random Forest (BRF).

2.3.3. Neural network

Neural Network has many advantages compared to other calculation methods, namely the ability to acquire knowledge even if there are disturbances and uncertainties. This is because the neural network can generalize abstraction and extraction of statistical properties from data. In addition, the neural network also can present capabilities in a flexible manner; a neural network can create its own representation through self-regulation or self-organizing skills. And there are many other advantages possessed by the neural network itself. The Figure 2 for architecture neural network:

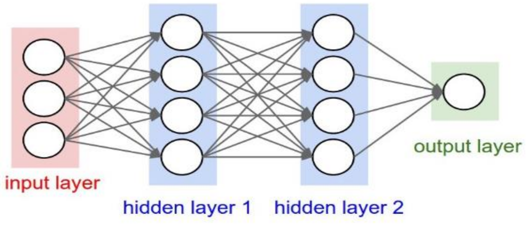

Figure 2. Architecture neural network

2.3.4. Support vector machine

The concept of SVM can be explained only as an attempt to find the best hyperplane 2 that functions as a separator of two classes in the input space. Figure 3 shows several patterns that are members of two classes: +1 and –1. Patterns that are joined in class 1 are represented as red (squares), while patterns in class +1 are represented as a yellow (circles). The calcification process is as shown:

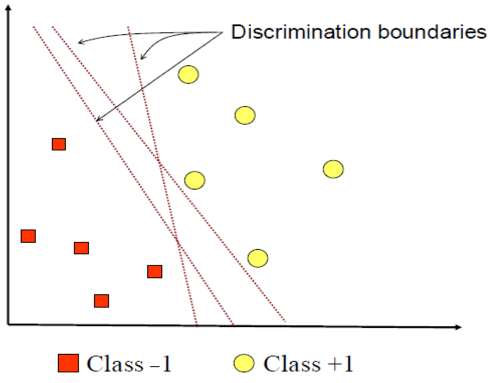

Figure 3. Hyperplane SVM

This problem can be solved by various computational techniques, including Lagrange Multiplier.

$$L(w,b,u) = \frac{1}{2}\|W\|^2 - \sum_{i-1}^{l} u_i[D_{ii}(W^T X_i - \gamma) - 1] \qquad (3)$$

αi is Lagrange multipliers, which are zero or positive (ai≥0). The optimal value of the equation can be calculated by minimizing L against w and b and maximizing L against αi.

2.3.5. K nearest neighbour

The K-Nearest Neighbor algorithm is a method that uses a supervised algorithm [20]. K-Nearest Neighbor includes instance-based learning groups. The K-Nearest Neighbor algorithm is simple, works based on the similarity of the test sample to the training sample (training sample) to determine the K-Nearest Neighbor [21] K-Nearest Neighbor is done by finding groups of k objects in the training data the closest (similar) to the object on new data or testing data [22]. K Nearest Neighbor is a simple classification technique, but it has good work results [23]. In general, to define the distance between two x and y objects, the Euclidean distance formula is used in the following equation:

$$dxy = \sqrt{\sum_{i=1}^{n}(Xi - Yi)^2} \qquad (4)$$

KNN has several advantages, namely toughness to training data that have a lot of noise and is effective when the training data are large. Meanwhile, the weakness of KNN is KNN need the value of the parameter k (number of closest neighbors), unclear distance-based training on what type of distance to use and which attributes should be used to get the best results, and computing costs are high because calculations are needed distance from each query instance in the whole training sample [15].

2.4. Evaluation metrics

Effective detection is the crux of our work; the wrong detection can prevent genuine packets from reaching their destinations. We want to calculate the accuracy of our detection mechanism for genuine and attack traffic and then compare it with other similar research that has reported accuracy. The performance of the classifiers is evaluated, and comparative analysis has been carried out. Classification accuracy is used as a primary performance measure for evaluating the classifiers and is measured as the ratio of the number of correctly classified instances in the test dataset and the total number of test cases. The performances of the trained models are evaluated based on the criteria of precision, recall, f-measure and accuracy using 10-fold cross validation [24].
formula for calculating accuracy is shown in (1)

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \text{ x } 100\% \qquad (5)$$

formula for calculating Recall is shown in (2)

$$\text{Recall} = \frac{TP}{TP+FN} \text{ x } 100\% \quad (6)$$

formula for calculating Precision is shown in (3)

$$\text{Precision} = \frac{TP}{TP+FN} \text{ x } 100\% \quad (7)$$

formula for calculating F-Measure is shown in (4)

$$\text{F} = 2*(\text{Precision} * \text{Recall})/(\text{Precision} * \text{Recall}) \quad (8)$$

3. RESULTS AND DISCUSSION

3.1. Feature selection

After analyzing DDoS dataset which has 25 features or features, Table 2 shows the analysis results are obtained. To detect DDoS attacks, the ideal features are packet delay, packet origin (from the node), destination packet (to node) and source IP Address. Of the 25 attributes contained in the dataset, only 4 attributes can be used to detect DDoS attacks, whereas the rest did not meet the criteria to be used as a tool for Classification in Machine Learning Techniques. To find out the value of R Square on each attribute can be explained in Table 3 as follows.

Table 2. Feature selection optimal results

Model	Variables Entered	Variables Removed	Method
1	PKT_DELAY	.	Forward (Criterion: Probability-of-F-to-enter <= ,050)
2	FROM_NODE	.	Forward (Criterion: Probability-of-F-to-enter <= ,050)
3	TO_NODE	.	Forward (Criterion: Probability-of-F-to-enter <= ,050)
4	SRC_ADD	.	Forward (Criterion: Probability-of-F-to-enter <= ,050)

a. Dependent Variable: PKT_CLASS

Table 3. Detect significant value using forward feature selection regression (ANOVA)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11,582	1	11,582	739,858	,000[a]
	Residual	58,060	3709	,016		
	Total	69,642	3710			
2	Regression	11,768	2	5,884	376,986	,000[b]
	Residual	57,874	3708	,016		
	Total	69,642	3710			
3	Regression	11,847	3	3,949	253,293	,000[c]
	Residual	57,795	3707	,016		
	Total	69,642	3710			
4	Regression	11,908	4	2,977	191,106	,000[d]
	Residual	57,733	3706	,016		
	Total	69,642	3710			

a. Predictors: (Constant), PKT_DELAY
b. Predictors: (Constant), PKT_DELAY, FROM_NODE
c. Predictors: (Constant), PKT_DELAY, FROM_NODE, TO_NODE
d. Predictors: (Constant), PKT_DELAY, FROM_NODE, TO_NODE, SRC_ADD
e. Dependent Variable: PKT_CLASS

It can be explained that each attribute has a sig value less than 0.05 (0,000 <0.05), meaning that the PKT_DELAY, FROM_NODE, TO_NODE, SRC_ADD attributes are very significant in detecting types of DDoS attacks such as Smurt Attack, UDP Flood, SQL Injection and HTTP Flood).

3.2. Algoritma machine learning

From the attributes that have been selected, training and testing are carried out on a dataset with 5 algorithms in accordance with the methods that have been determined can be seen in the Table 4 and Figure 4-7 as follows:

Table 4. Dataset detail

Dataset	Number Dataset DDoS
Training	5899
Testing 30%	1770

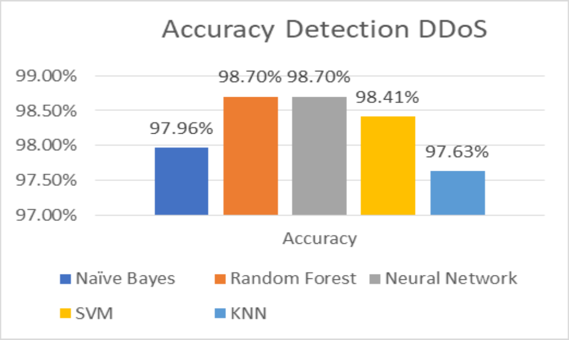

Figure 4. Accuracy detection graphic

The highest level of accuracy for detecting DDoS attacks is using the Random Forest algorithm and the Neural Network of 98.70%.

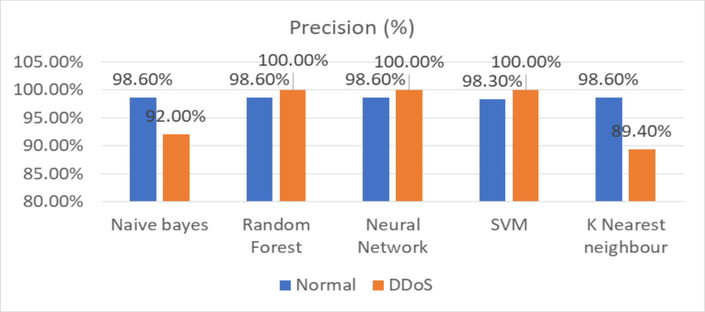

Figure 5. Precision detection graphic

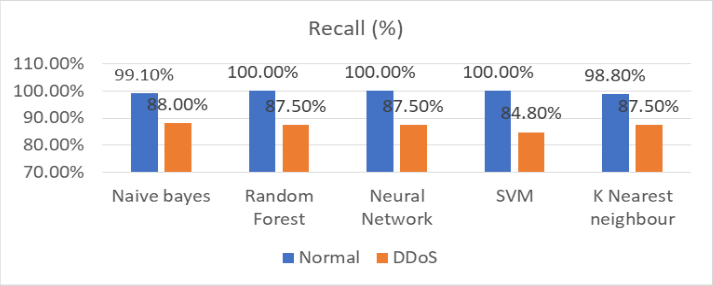

Figure 6. Recall detection graphic

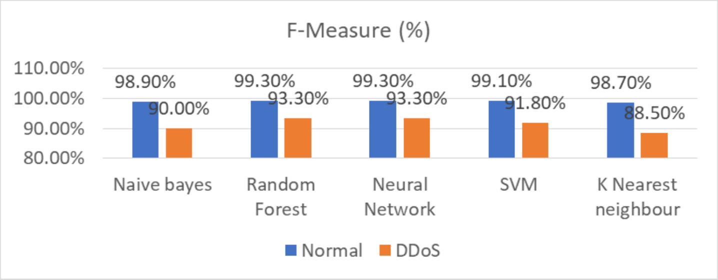

Figure 7. F-Measure detection graphic

4. CONCLUSION

In this paper, we put together a new dataset that covers the types of modern attacks, which were not used in previous studies. The dataset contains 25 features and five classes. Attacks are carried out directly to the target server and capture packet data using a high-trust Wireshark application because of its ability to produce valid results that reflect the real environment. Collected data has been recorded for various types of attacks that target the network Application layer. From the datasets, there is a training and testing of data using five techniques classification: Neural Network, Naïve Bayes and Random Forest, KNN, and Support Vector Machine (SVM), datasets processed have different percentages, with the aim of facilitating in classifying. From this study it can be concluded that from the five classification techniques used, the Forest random classification technique achieved the highest level of accuracy (98.70%) with a Weighted Average of 98.4%. This means that the technique is able to detect DDoS attacks accurately on the application that will be developed

ACKNOWLEDGMENTS

We would like to thank all to the University of Putera Batam Indonesian and Universiti Tun Hussein Onn Malaysia, who have provided practice space in this research.

REFERENCES

[1] Abdul Kadir, *Introduction to the Revised Edition Information System*. 2014.

[2] M. M. Irsyad, “Analysis System Anomaly Traffic Detection with Comparing The Differences of Triangle-Area-Map Features for Anomaly Type Identification Mujp,” *Telkom Univ.*, vol. 2, no. 1, pp. 254–263, 2015.

[3] K. Kato and V. Klyuev, “An Intelligent DDoS Attack Detection System Using Packet Analysis and Support Vector Machine,” *Int. J. Intell. Comput. Res.*, vol. 5, no. 3, pp. 464–471, 2014.

[4] K. Ramadhani, M. Yusuf, and H. E. Wahanani, “Cusum-Based Traffic Change Anomaly,” 2014.

[5] A. W. Muhammad and I. Riadi, “Detection of DDoS Attacks Using Neural Network with Fixed Moving Average Window Function,” vol. 1, no. 3, pp. 115–122, 2017.

[6] A. Oza, “HTTP Attack Detection using N-gram Analysis,” 2013.

[7] T. P. Thwe Thwe Oo, “A statistical approach to classify and identify DDoS attacks using UCLA dataset,” *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 2, no. 5, p. 1766, 2013.

[8] J. David and C. Thomas, “DDoS Attack Detection using Fast Entropy Approach on Flow-Based Network Traffic,” *Procedia - Procedia Comput. Sci.*, vol. 50, pp. 30–36, 2015.

[9] M. Alkasassbeh, A. B. A. Hassanat, and G. Al-naymat, “Detecting Distributed Denial of Service Attacks Using Data Mining Techniques,” vol. 7, no. 1, pp. 436–445, 2016.

[10] B. A. Pratomo and R. M. Ijtihadie, “Sistem Deteksi Intrusi Menggunakan N-Gram Dan Cosine Similarity,” *JUTI J. Ilm. Teknol. Inf.*, vol. 14, no. 1, p. 108, 2016.

[11] S. Sridharan, “Defeating n-gram Scores for HTTP Attack Detection,” 2016.

[12] A. Oza, K. Ross, R. M. Low, and M. Stamp, “HTTP Attack Detection using N-gram Analysis.pdf,” *Comput. {&} Secur.*, vol. 45, pp. 242–254, 2014.

[13] I. Riadi, A. W. Muhammad, and Sunardi, “Neural network-based DDoS detection regarding hidden layer variation,” *J. Theor. Appl. Inf. Technol.*, vol. 95, no. 15, pp. 3684–3691, 2017.

[14] B. A. Tama and K. H. Rhee, “Data mining techniques in DoS / DDoS attack detection : A literature review Data Mining Techniques in DoS / DDoS Attack Detection : A Literature Review,” no. August 2015, 2017.

[15] U. S. Utara, U. S. Utara, and U. S. Utara, “Botnet Detection Using the K-Nearest Neighbor Algorithm,” 2018.

[16] M. Alkasassbeh, G. Al-Naymat, A. B.A, and M. Almseidin, “Detecting Distributed Denial of Service Attacks Using Data Mining Techniques,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 1, 2016.

[17] C. Max-dependency, “Feature Selection Based on Mutual Information :” vol. 27, no. 8, pp. 1226–1238, 2005.

[18] E. Manalu, F. A. Sianturi, and M. R. Manalu, “Application of Naive Bayes Algorithm To Predict The Production Amount Based On Inventory Data And the number of ordering on cv. Papadan Mama Pastries,” vol. 1, no. 2, 2017.

[19] T. Mahjabin, Y. Xiao, G. Sun, and W. Jiang, “A survey of a distributed denial-of-service attack, prevention, and mitigation techniques,” vol. 13, no. 139, 2017.

[20] M. 2006. D. M. C. and T. N. Y. M. K. P. Han, J., & Kamber, *Data mining Concept and Techniques. New York*. 2006.

[21] Siringoringo, “Comparative Analysis of Cluster Process Using K-Means Clustering and K-Nearest Neighbor in Diabetes Mellitus,” 2016.

[22] K. Kepemilikan and K. Bemotor, “Application of k-nearest neighbor algorithm for determining credit risk of motorized vehicle ownership,” vol. 1, no. 1, pp. 65–76, 2013.

[23] U. S. Utara, “Universitas Sumatera Utara,” 2015.

[24] L. M. Shi, A. Mustapha, Y. Mazwin, and M. Hassim, “Predicting fatalities among shark attacks: comparison of classifiers,” vol. 8, no. 4, pp. 360–366, 2019.