

C. Feature Selection

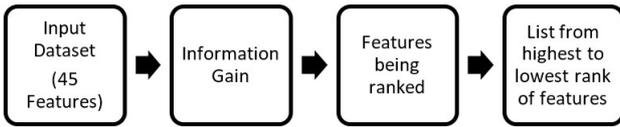


Fig. 1 Feature Selection Process

Based on Figure 1, there are 45 features available in this dataset. To get the best features to detect the DDoS attack, this dataset will go through the features selection step. The 45 features available will be selected by a filter method called Information Gain. This technique will select the features based on information gained and item frequency. This method's fitness function has been enhanced to carefully consider the weight, text, and vector similarity dimensions [22].

D. Pre-processing



Fig 2 Pre-processing Process

Based on Figure 2, after the features have been listed and ranked in the features selection step, important and less important features have been successfully identified. It is the time to use the Data Reduction method to select only important features that will be used in the classification phase. The rest of the features that have not been selected will be removed from the dataset. This research will select 10 features that are most important and helpful in detecting DDoS attacks. Thus, the result of selected features will be shown in the next chapter.

E. Classification

The first classifier that will be observed in this research is Artificial Neural Network. The process of the classification will be shown in Figure 3.

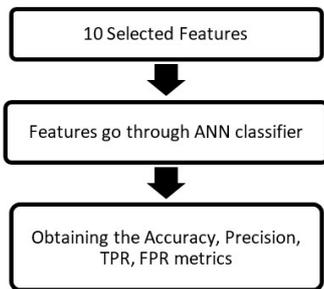


Fig. 3 Classification Process for ANN

Based on Figure 4, 10 features that have been selected in the pre-processing phase will go through the classification process. The data will be run in the ANN classifier in the WEKA tool. Artificial neural networks are divided into layers based on the number of parallel computing processes they perform [23]. Each of the number of inputs is multiplied by an initially established weight for each processor in a layer, resulting in the internal value of the operation. After the

process finishes being run, the WEKA tool will generate the parameter evaluation metrics which are Accuracy, Precision, TPR and FPR. The second classifier that will be observed in this research is Naïve Bayes.

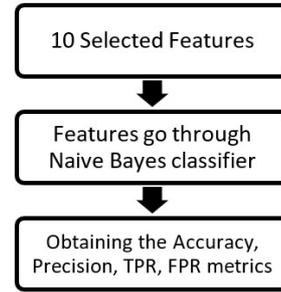


Fig. 4 Classification Process for Naïve Bayes

Based on Figure 4, 10 features that have been selected in the pre-processing phase will go through the classification process. At this phase, the data will be run in the Naïve Bayes classifier in the WEKA tool. This Naïve Bayes classifier requires quite less training data and is highly extensible [24]. After the process finishes being run, the WEKA tool will generate the parameter evaluation metrics which are Accuracy, Precision, TPR and FPR. The result of this process will be shown in the next chapter.

The third classifier that will be observed in this research is the Decision Table. The process of the classification will be shown in Figure 5.

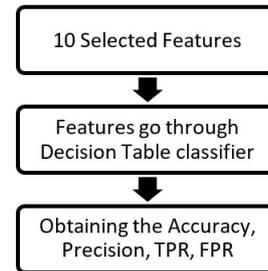


Fig. 5 Classification Process for Decision Table

Based on Figure 5, 10 features that have been selected in the pre-processing phase will go through the classification process. The data will be run in the Decision Table classifier in the WEKA tool. In a decision table, conditions are usually expressed as True (T) or False (F). Each column in the table corresponds to a rule in the business logic that describes the unique combination of circumstances that will result in the actions [25]. After the process finishes being run, the WEKA tool will generate the parameter evaluation metrics which are Accuracy, Precision, TPR and FPR. The result of this process will be shown in the next chapter.

III. RESULT AND DISCUSSION

A. Dataset Result

The dataset downloaded is UNSW-NB 15 which is the same dataset being used in [13]. The dataset is then saved into the CSV file to run in WEKA for the experiment process. To obtain accurate results in this study, the UNSW-NB 15 Dataset was divided into two sets of files by using the ratio scale measurement method. This dataset is divided using a

ratio of 70:30 which is 70% for the training set and 30% for the testing set. The training set contains 175341 records while the testing set has 82332 records. The distribution of UNSW-NB 15 datasets in a ratio of 70:30.

In this dataset phase, the ratio used was 70:30. Training set has 175341 instances (70%) while the testing set has 82332 instances (30%). These training and testing sets are very important to run the experimental process in this research.

B. Features Selection and Pre-processing Result

In this process, all 45 features will be listed and ranked. The features will be ranked from the most important features to the less important based on score. This process is very important to know where features are important and not important in detecting the DDoS attack. After that, 10 highest features from the 45 ranked features will be selected as the subset of the features by looking at the scores in the pre-processing phase. 10 features that have been selected were then scaled down and transferred into another CSV file. 10 features that have been selected will be used for the classification part. Work in [13] also applies this dataset to perform the classification. The subset of features they are using are different from this research.

After running the classification, the accuracy, true positive rate and false positive rate for the selected features has been obtained for each classifier. Table 1 and Table 2 will show the comparison of classification results between the selected features.

TABLE I
COMPARISON OF RESULT BETWEEN 10 SELECTED FEATURES AND FEATURES IN [13] USING ANN CLASSIFIER

Features	Accuracy	True Positive Rate	False Positive Rate
Selected Features	84.66 %	0.847	0.020

TABLE II
COMPARISON OF RESULT BETWEEN 10 SELECTED FEATURES AND FEATURES IN [13] USING NAÏVE BAYES CLASSIFIER

Features	Accuracy	True Positive Rate	False Positive Rate
Selected Features	87.66 %	0.887	0.008

This research only compared results of ANN and Naïve Bayes because these classifiers are being used in this research and [13] research. Based on Table 1, the results for ANN classifiers for this research's features have the Accuracy (84.66%), TPR (0.847), and FPR (0.020). For Table 2, Naïve Bayes classifier, classification for proposed features resulted in Accuracy (87.66 %), TPR (0.887), and FPR (0.008).

This research mentioned that the same dataset was used by research in [13]. Tuan [13] also uses the UNSW-NB 15 dataset in the detection of DDoS attacks. But [13] are using different features in this dataset to make the experiment of DDoS attack detection.

C. Classification Result

Table 3 shows the classification using 45 original features from the dataset with several parameters namely Accuracy,

Precision, TPR and FPR. For Accuracy, the Decision Table score was the best at 84.54% followed by ANN 83.98%, and Naïve Bayes 81.97%. For Precision, the Decision Table is highest with 0.897 followed by Naïve Bayes with a score of 0.894 and ANN with 0.890. For True Positive Rate, the highest value is the best while for False Positive, the lowest value is the best. For TPR, Decision Table got 0.865 and ANN got 0.840, while Naïve Bayes got 0.820. For False Positive Rate, ANN is the highest with 0.010, Naïve Bayes got 0.011, and Decision Table got 0.019.

TABLE III
CLASSIFICATION USING ANN, NAÏVE BAYES, AND DECISION TABLE ON 45 ORIGINAL FEATURES

Classifier	Accuracy	Precision	True Positive Rate	False Positive Rate
ANN	83.98 %	0.890	0.840	0.010
Naïve Bayes	81.97 %	0.894	0.820	0.011
Decision Table	84.54 %	0.897	0.865	0.019

Table 4 shows the classification using 10 selected features with several parameters namely Accuracy, Precision, True Positive and False Positive. For Accuracy, the Decision Table score is the best at 88.43% followed by Naïve Bayes 87.74%, and Artificial Neural Network (ANN) 84.66%. For Precision, ANN result was 0.912 followed by Naïve Bayes with a score of 0.905 and Decision Table with 0.896. For True Positive Rate, the highest value is the best while for False Positive, the lowest value is the best. For TPR, Naïve Bayes got 0.887, ANN 0.847 and Decision Table got 0.884. For False Positive Rate, Decision Table got 0.021, ANN got 0.020, and Naïve Bayes got 0.008.

TABLE IV
CLASSIFICATION USING ANN, NAÏVE BAYES, AND DECISION TABLE ON 10 SELECTED FEATURES

Classifier	Accuracy	Precision	True Positive Rate	False Positive Rate
ANN	84.66 %	0.912	0.847	0.020
Naïve Bayes	87.74 %	0.905	0.887	0.008
Decision Table	88.43 %	0.896	0.884	0.021

There is an experiment to find the best classifier in the classification process. The classifiers that have been used are ANN, Naïve Bayes and Decision Table whereas the Accuracy is based on percentage values (1 to 100) and for validation of all the methods this research used a full training and testing set for UNSW-NB15 dataset. Based on the experimental analysis, this research observed that Decision Table and Naïve Bayes classifiers are the best at detecting and differentiating the DDoS attack and normal network traffic with an accuracy of 88.43% for Decision Table and 87.74% for Naïve Bayes. These algorithms make use of Information Gain for selecting the best features in detecting DDoS attacks. For the Decision Table, the accuracy is (88.43%), Precision (0.896), True Positive Rate (0.884) and False Positive Rate (0.021). For Naïve Bayes, the accuracy is (87.84%), Precision (0.905), True Positive Rate (0.887) and False Positive Rate

(0.008). Thus, as stated in the problem statement, there is previous research that does not apply Decision Table in their experiments. Thus, this research proved that Decision Table is also a good classifier in detecting DDoS because it got the highest score (88.43%) in our DDoS attack detection experiment.

In this phase, there is also an extra experiment which is the comparison of classification results between the 45 original features from the dataset and 10 selected features for this research. Table 4 shows the classification using 45 original features from the UNSW-NB 15 dataset while Table 5 shows the classification on a testing set using 10 selected features for this research. Based on those tables, we can see that after we do the features selection and pre-processing step to reduce the number of features, the classification accuracy of each classifier has been increased. For 45 original features, the accuracy of ANN is 83.98%, Naïve Bayes 81.97% and Decision Table is 84.54%. For 10 selected features, accuracy of ANN is 84.66%, Naïve Bayes 87.74% and Decision Table is 88.43%. This shows that the result would be increased when we apply the Information Gain and reduce the number of features in the pre-processing phase.

IV. CONCLUSION

In conclusion, the framework of machine learning techniques for DDoS attack detection has been studied, developed and evaluated with Accuracy, Precision, TPR, and FPR by using ANN, Naïve Bayes and Decision Table until it achieved this research's objectives. The evaluation was done on the UNSW-NB 15 dataset, which is the well-known dataset for DDoS attack detection nowadays. To recap the problem statement, the first problem is DDoS attacks have increased in recent years although there are many researchers who do research in this problem. Hence, the first problem has been overcome by this research has developed an enhanced framework of DDoS attack detection by using machine learning techniques. The framework would start with Information Gain in feature selection, Data Reduction in pre-processing, using ANN, Naïve Bayes and Decision Table as classifiers, and Accuracy, Precision, TPR and FPR for the metrics. The second problem is there are many researches that do not include the Decision Table classifier in their experimental analysis. Hence, this research proved that Decision Table is also a good classifier in detecting DDoS because it got the highest score (88.43%) in our DDoS attack detection experiment. In computer security and other related fields, this validation is important. These problems statement has been overcome by doing every experiment in this research. In the experimental analysis, we observe that Naïve Bayes and Decision Table classifiers are the best at detecting and differentiating the DDoS attack and normal network traffic in terms of Accuracy, Precision, True Positive Rate and False Positive Rate. This research also provided an enhanced DDoS detection framework by doing the feature selection approach with Information Gain that generates a better classification result from previous research by adding as discussed in the discussion section.

The proposed framework has several drawbacks. It was just observed on a single dataset which is UNSW-NB 15. There several more well-known datasets may be used to evaluate the machine learning algorithms' efficiency. Moreover, only

Distributed Denial of Service (DDoS) attacks were investigated and detected in this research. Therefore, multiple more attacks might be investigated and analyzed in the future.

ACKNOWLEDGMENT

This research was supported by the Ministry of Higher Education (MOHE) through Fundamental Research Grant Scheme (FRGS/1/2018/ICT04/UTHM/03/4).

REFERENCES

- [1] M. Zekri, S. El Kafhali, N. Aboutabit, and Y. Saadi, "DDoS attack detection using machine learning techniques in cloud computing environments," in *Proceedings of 2017 International Conference of Cloud Computing Technologies and Applications, CloudTech 2017*, 2018, vol. 2018-January, pp. 1–7.
- [2] T. Mahjabin, Y. Xiao, G. Sun, and W. Jiang, "A survey of distributed denial-of-service attack, prevention, and mitigation techniques," *Int. J. Distrib. Sens. Networks*, vol. 13, no. 12, Dec. 2017.
- [3] D. Albertivan, H. Limantara, R. A. Rachmadiati, A. W. Pamungkas, and N. Surantha, "IT risk identification and evaluation: A case study on XYZ University," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 9, no. 1, pp. 250–257, 2019.
- [4] A. K. Hakim, M. Abdurrohman, and F. A. Yulianto, "Improving DDoS detection accuracy using Six-Sigma in SDN environment," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 8, no. 2, pp. 365–370, 2018.
- [5] S. Suroto, "A Review of Defense Against Slow HTTP Attack," *JOIV Int. J. Informatics Vis.*, vol. 1, no. 4, pp. 127–134, Nov. 2017.
- [6] S. Cook, "Comparitech," 2020. [Online]. Available: <https://www.comparitech.com/blog/information-security/ddos-statistics-facts/>. [Accessed: 24-Oct-2020].
- [7] B. B. Gupta, R. C. Joshi, and M. Misra, "Dynamic and Auto Responsive Solution for Distributed Denial-of-Service Attacks Detection in ISP Network," Apr. 2012.
- [8] A. Srivastava, B. B. Gupta, A. Tyagi, A. Sharma, and A. Mishra, "A recent survey on DDoS attacks and defense mechanisms," in *Communications in Computer and Information Science*, 2011, vol. 203 CCIS, pp. 570–580.
- [9] F. M. Nur Aini Zafrah Bt Abdul Kamal, "Botnet Malware Network Intrusion Detection Using Machine Learning Universiti Tun Hussein Onn Malaysia," 2019.
- [10] R. Vishwakarma, "A HoneyPot with Machine Learning based Detection Framework for defending IoT based Botnet DDoS Attacks," 2019.
- [11] R. Vishwakarma and A. K. Jain, "A survey of DDoS attacking techniques and defence mechanisms in the IoT network," *Telecommunication Systems*, vol. 73, no. 1, Springer, pp. 3–25, 2020.
- [12] A. Chowdhury, "Recent cyber security attacks and their mitigation approaches - An overview," in *Communications in Computer and Information Science*, 2016, vol. 651, pp. 54–65.
- [13] T. A. Tuan, H. V. Long, L. H. Son, R. Kumar, I. Priyadarshini, and N. T. K. Son, "Performance evaluation of Botnet DDoS attack detection using machine learning," *Evol. Intell.*, vol. 13, no. 2, pp. 283–294, Jun. 2020.
- [14] M. Barati, A. Abdullah, N. I. Udzir, R. Mahmud, and N. Mustapha, "Distributed Denial of Service detection using hybrid machine learning technique," in *Proceedings - 2014 International Symposium on Biometrics and Security Technologies, ISBAST 2014*, 2015, pp. 268–273.
- [15] B. Zhou, J. Li, J. Wu, S. Guo, Y. Gu, and Z. Li, "Machine-learning-based online distributed denial-of-service attack detection using spark streaming," in *IEEE International Conference on Communications*, 2018, vol. 2018-May.
- [16] S. R. Kalmegh, "Comparative Analysis of the WEKA Classifiers Rules Conjunctiverule & Decisiontable on Indian News Dataset by Using Different Test Mode," 2018.
- [17] P. Kaur, M. Kumar, and A. Bhandari, "A review of detection approaches for distributed denial of service attacks," *Syst. Sci. Control Eng.*, vol. 5, no. 1, pp. 301–320, Jan. 2017.
- [18] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, Jul. 2018.

- [19] Sui, "Information Gain Feature Selection Based on Feature Interactions," 2013.
- [20] L. Koesten, E. Simperl, T. Blount, E. Kacprzak, and J. Tennison, "Everything you always wanted to know about a dataset: Studies in data summarisation," *Int. J. Hum. Comput. Stud.*, vol. 135, p. 102367, Mar. 2020.
- [21] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *2015 Military Communications and Information Systems Conference, MilCIS 2015 - Proceedings*, 2015.
- [22] S. Lei, "A feature selection method based on information gain and genetic algorithm," in *Proceedings - 2012 International Conference on Computer Science and Electronics Engineering, ICCSEE 2012*, 2012, vol. 2, pp. 355–358.
- [23] A. F. Kuri-Morales, "The best neural network architecture," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014, vol. 8857, pp. 72–84.
- [24] Z. K. Maseer, R. Yusof, N. Bahaman, S. A. Mostafa, and C. F. M. Foozy, "Benchmarking of Machine Learning for Anomaly Based Intrusion Detection Systems in the CICIDS2017 Dataset," *IEEE Access*, vol. 9, pp. 22351–22370, 2021.
- [25] G. Witt, "A brief history of rules," in *Writing Effective Business Rules*, Elsevier, 2012, pp. 25–63.