ICIC Express Letters
Part B: Applications
Volume **13**, Number **4**, April **2022**

ICIC International ©2022 ISSN 2185-2766
pp. 347–354

PORTMAP DDOS ATTACK DETECTION USING FEATURE RANK AND MACHINE LEARNING ALGORITHMS

Yuna Sugianela and Tohari Ahmad*

Department of Informatics
Institut Teknologi Sepuluh Nopember
Kampus ITS, Surabaya, Jawa Timur 60111, Indonesia
yuna.sugianela13@mhs.if.its.ac.id; *Corresponding author: tohari@if.its.ac.id

Received May 2021; accepted August 2021

Abstract. *The era of big data, which is coming with a complicated and big scope of data, has caused the increase of the possibility of network attack. One of those possible attacks is DDoS or Distributed Denial of Service. It is a type of attack that floods the network traffics, and it usually is implemented in the upper layers of the network protocol. DDoS occurs like a highway blocked by traffic jams so that traffic flow does not arrive at the desired destination. Some research generates datasets of network attacks, especially on this DDoS. They analyze the taxonomy of attacks or determine important factors that affect the corresponding attack. The method for detecting DDoS is usually done by an Intrusion Detection System (IDS) using classification and clustering methods. Machine learning has been widely used to make IDS optimal. Despite the fact that a machine learning algorithm has good adaptability to detect the attack, it needs time for processing the dataset with high dimensional data, for example, 80 features. In this paper, we propose the feature selection using feature rank and the detection using some machine learning algorithms to balance the dimensionality of data and the accuracy. We focus on detecting the PortMap DDoS attack as the reflection-based DDoS. The proposed method reaches the most effective result in 99.937% of accuracy and consumes 0.04 seconds from the Chi-square attribute evaluation with stopping criteria of 7000 with the k-NN classification method.*

Keywords: Network security, Features selection, Network infrastructure, Classification, Data protection

1. **Introduction.** For years, Information and Communication Technology (ICT) has grown fast and has played an essential role in many aspects of life. People can connect easily because of computer networks. Moreover, a large amount of data is processed that the public generates about five exabytes of data per two days [1]. In this big data era, which is a period with a complicated and big scope of data, the possibility of network attack has increased [2].

In line with the increased use of data, various threats have emerged, such as the Distributed Denial of Service (DDoS) attack. It has been one of the main concerns of computer network security since the last decades [3]. DDoS works by flooding the network traffic and is likely to be implemented in some protocol layers, such as network, transport, and application [4]. DDoS prevents the legitimate packet data from moving by blocking the network, making them unable to reach the destination. Furthermore, we can use data to analyze the behavior of computer network attacks like DDoS and determine the method to anticipate it. Some research generates the dataset of DDoS, analyzes the attack taxonomy, and finds important factors that affect the attack. Besides, some research is to detect whether the incoming packet is an attack or not.

DOI: 10.24507/icicelb.13.04.347

The taxonomies of DDoS attacks have been widely studied. In this research, we refer to Sharafaldin et al. [4] who analyze the new attack, as depicted in Figure 1. It may be performed using either a UDP (User Datagram Protocol)-based protocol, TCP (Transmission Control Protocol)-based protocol, or both at the application layer. There are two types of DDoS attacks: reflection-based and exploitation-based attacks. Both of them hide the identity of the attacker using a third-party component. There are reflector servers that receive packets whose IP source address is located at the victim's target IP number to flood it with the replying packets.

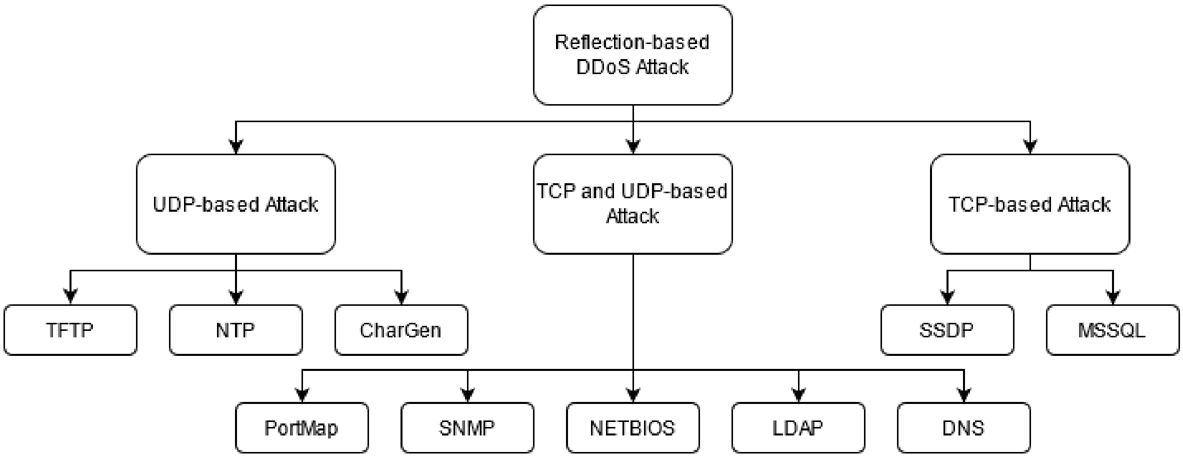

FIGURE 1. The classification of reflection-based DDoS attack, based on [4]

The difference between reflection-based and exploitation-based attacks is that the former can also work through a combination of TCP and UDP [4]. The reflection-based DDoS begins when the attacker sends a query to a target's IP number. The destination may be any machine on the Internet that carries out reflective UDP services. Packets from the attacker are faked to look like it is not coming from the victim. They may take an advanced tool to transfer dangerous queries at high speed massively, which is then responded to by the victims. One of the reflection-based DDoS attacks is PortMapper which is also known as PortMap and RPCbind. This attack produces 7-28 times amplification in bandwidth. PortMap runs on services that are used to direct clients to the correct port number, namely on UDP or TCP port 111, so that they can communicate with the requested RPC service [5].

Currently, DDoS attacks are a concern to investigate. This includes determining important features to recognize the attack faster and easier [1], and grouping like clustering and classification for detection [3, 6, 7, 8]. The machine learning algorithm, which has been broadly used for optimally detecting intrusion, is a computational science that follows human intelligence that is currently developing. Models of statistics are used for the machine learning method to discover patterns in data with large volumes [9], for example, hidden Markov models, Neural Networks (NN), Support Vector Machines (SVM), and fuzzy logic [10]. In 2019, Sharafaldin et al. [4] proposed a new taxonomy of DDoS attacks and generated a new dataset, namely CICDDoS2019 [11].

In further research, Aziz and Ahmad [12] used a cluster analysis-based approach to select the best feature clusters. For the classification, they take SVM, Naive Bayes, and J-48 methods. They can achieve the highest accuracy at 99.842%. In [13], we used the Pearson's correlation method to rank the best feature of an intrusion detection system. In that study, we can reduce the running time and reach 99.36% of accuracy.

In this research, we focus on detecting a PortMap attack as the reflection-based DDoS. With specific data types, we hope that the detection results can be more detailed. Furthermore, we evaluate the available features and predict which contribute to the classification.

In the previous study, Deka et al. [1] proposed feature selection in DDoS attacks using parallel computing and parallel ranker algorithms, whose detection accuracy is 97%. In our research, the accuracy reaches 99.937% with a processing time of 0.04 seconds.

This paper is divided into four parts. The first is the introduction, which contains the background and brief description of the research related to this study. In the second section, we explain the methods that we use in this research. The next section explains the dataset, the environment, and the experimental result. The last part is the conclusion of the research.

2. **Proposed Method.** In this study, we focus on the detection of the PortMap as the reflection-based DDoS attack, whose process is given in Figure 2. In the first step, we select 10% of the dataset for the experiment. Next, we select the feature to reduce the dimensional data using feature rank. The last step is classification using the machine learning algorithm. In this research, we use Random Forest (RF), k-Nearest Neighbor (k-NN), and J-48 methods.

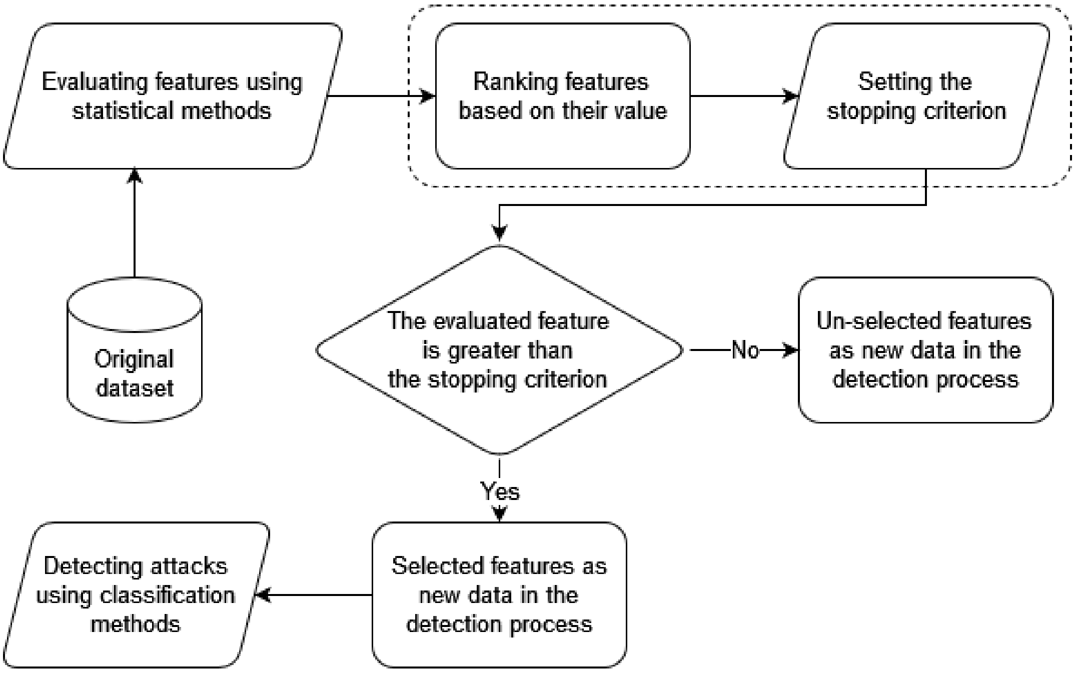
Evaluating features using statistical methods
Ranking features based on their value
Setting the stopping criterion
Original dataset
The evaluated feature is greater than the stopping criterion
No
Un-selected features as new data in the detection process
Yes
Detecting attacks using classification methods
Selected features as new data in the detection process

FIGURE 2. The focus of the proposed method

2.1. **Feature selection using ranked feature.** Feature selection is a process to reduce the dimension of N features of data into M features, where M is always less than N. The reduction process is intended to speed up the learning algorithm by still obtaining a relatively good detection accuracy and improve the comprehensibility of learning results. Feature selection is a search problem according to some measurement criteria. There are two techniques in the feature evaluation domain: subset evaluation and individual attribute evaluation [14].

1) Feature evaluation using statistical method. In this research, we use the individual attribute evaluation to evaluate the degree of relevance of features in the PortMap DDoS dataset. The statistical computation is applied to getting the value of the feature's evaluation. Here, we compare three statistical methods to get the feature's evaluation's value: Pearson's correlation, information gain attribute evaluation, and Chi-square.

a) Pearson's correlation. In this evaluation, we calculate the correlation coefficient between each feature and output variables, and then select the features whose correlation is moderate-to-high positive or negative (between -1 and 1). On the other hand, features with a low correlation or close to zero value are dropped. For achieving this purpose, Pearson's correlation coefficient ρ_i in (1) is implemented, where X is the feature, and Y is the target. The $cov(X_i, Y)$ is the covariance, and σ is the standard deviation.

$$\rho_i = \frac{cov(X_i, Y)}{\sigma(X_i)\sigma_Y} \tag{1}$$

b) Information gain attribute evaluation. The evaluation method using information gain is done by calculating the gain value of each feature, which is obtained from the entropy value before separation subtracted by the entropy value after separation. Entropy is generally used to measure the uncertainty of a set of features from a dataset. This measurement is considered as a measure of uncertainty, where the higher the entropy, the higher the uncertainty [15].

c) Chi-square. The purpose of the Chi-square evaluation is to determine the correlation between variables. The type of data used in the Chi-square evaluation must be in the form of nominal or ordinal, periodic frequency data, or one of the nominal or ordinal data. This evaluation method is part of the non-parametric statistical analysis. The Chi-square value is calculated using Chi-square-metric in [16].

2) Rank the feature by the value of feature evaluation. In each feature evaluation step, we get the evaluation value of every feature. The range of value in every evaluation method is different, which is then sorted ascendingly. Some features are selected from the best-ranked ones whose value is greater than the specified stopping criterion.

3) Set the stopping criterion to select some best features. We propose the stopping criterion value to select the best features, which is different for every statistical method in evaluating features. In our previous research [13], the best stopping criterion value in Pearson's correlation method for IDS dataset [17] is 0.2, which reaches the best accuracy in the random forest method. In that research, selecting features reduces the processing time. We set stopping criterion value (s) for every feature evaluation method and evaluate them to get the best detection accuracy. Features having the value of evaluation greater than s are selected as a new dataset for the detection step.

2.2. **Classification using machine learning.** We get a new dataset from the previous selection step; each of its rows is detected from the PortMap attack or benign using classification. In this research, we take some machine learning methods in the classification step: Random Forest (RF), k-Nearest Neighbor (k-NN), and J-48. The parameter for the classification is provided in Table 1.

TABLE 1. Classification methods and their parameter

Classification method	Parameter
RF	The bag size percent and the batch size as 100, whose seed is 1
k-NN	Linear NN search, Euclidean distance for distance function
J-48	Binary splits with false value, the value of confidence factor is 0.25, and the seed is 1

3. **Main Results.** In this study, we use the CICDDoS2019 dataset, which is provided by the Canadian Institute for Cybersecurity, University of New Brunswick (UNB) [11]. The dataset contains both reflection-based and exploitation-based DDoS attacks. We take that first type of attack dataset, especially the PortMap attack. The original PortMap attack data has a total of 191695 lines, 80 features, and two classes, namely benign and attack. Here, we use 10% of all PortMap data, comprising 473 and 18696 records of benign and attack data, respectively.

The proposed method is implemented in the Weka 3.6.13, running on Ubuntu 16.04.6 LTS with 16 GB RAM. The accuracy (Acc) is determined by (2), whether those in the PortMap dataset is classified correctly as benign or attack. We also evaluate the performance of the proposed method using a calculation of the time consumption in the detection process.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

TP or True Positive is a condition when the attack data are correctly detected. FP or False Positive is a condition when data are benign but is detected as an attack. TN or True Negative is when benign data are correctly classified as benign data. Finally, False Negative (FN) is a condition when the attack data are recognized as benign. We measure the detection time to find which statistical method properly ranks the features, so the number of features reduces efficiently. To evaluate the effect of the feature reduction, its accuracy is also measured.

After obtaining the ranked features, we determine the value of the appropriate stopping criterion for each statistical method, as shown in Table 2. This ranking indicates important features that affect the detection results. In this table, we show some important features and those close to the stopping criterion.

Table 2. Stopping criteria for ranked features

Statistical method	Stopping criteria		
Pearson's correlation	0.1	0.2	0.3
Information gain	0.08	0.09	0.1
Chi-square	5000	7000	9000

In the information gain attribute evaluation, values of the evaluation step are in the range from 0 to 0.145591. There are eleven features that have the 0 value. Based on this result, we select three of them and evaluate them to get the best stopping criterion in the features reduction step. From the first value of this stopping criterion, which is 0.1, we get 20 new selected features; and from 0.09, we have 25 new selected features. The third value is 0.08, which generates 31 features.

Using Pearson's correlation, the obtained value is from 0 to 0.61472; twelve features generate the 0 value. In this evaluation, we also select three values to get the best stopping criterion. According to our previous research [13], we set 0.1, 0.2, and 0.3, which lead to 53, 42, and 21 selected features, respectively. Regarding the Chi-square, the value ranges from 0 to 13721.19637. Here, eleven features have the 0 value. By specifying 5000, 7000, and 9000 to be the first stopping criterion, we get 39, 30, and 26 new features.

As summarized in Table 2, it is also found that the higher the stopping criterion, the smaller number of features we get. Furthermore, the number of selected features affects the time consumption, as shown in Figures 3, 4, and 5. It can be inferred that fewer features tend to take less processing time. By using reduced features, we also evaluate the accuracy of detection whose results are provided in Table 3. It depicts that, in general, a higher number of features leads to higher accuracy. For all classifiers implemented with information gain attribute evaluation, the best accuracy is always obtained with the

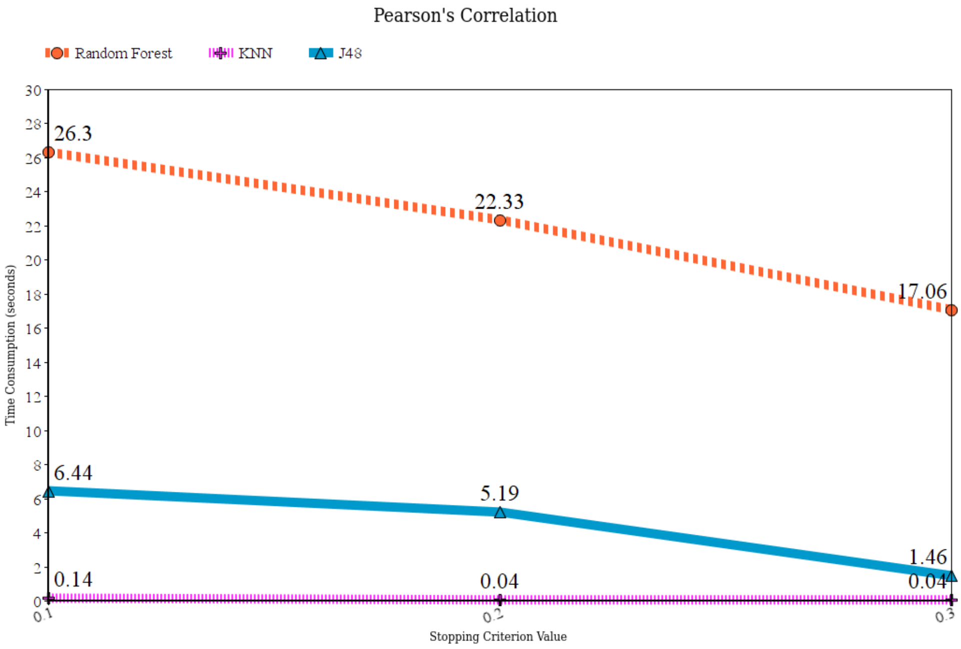

FIGURE 3. Execution time using Pearson's correlation

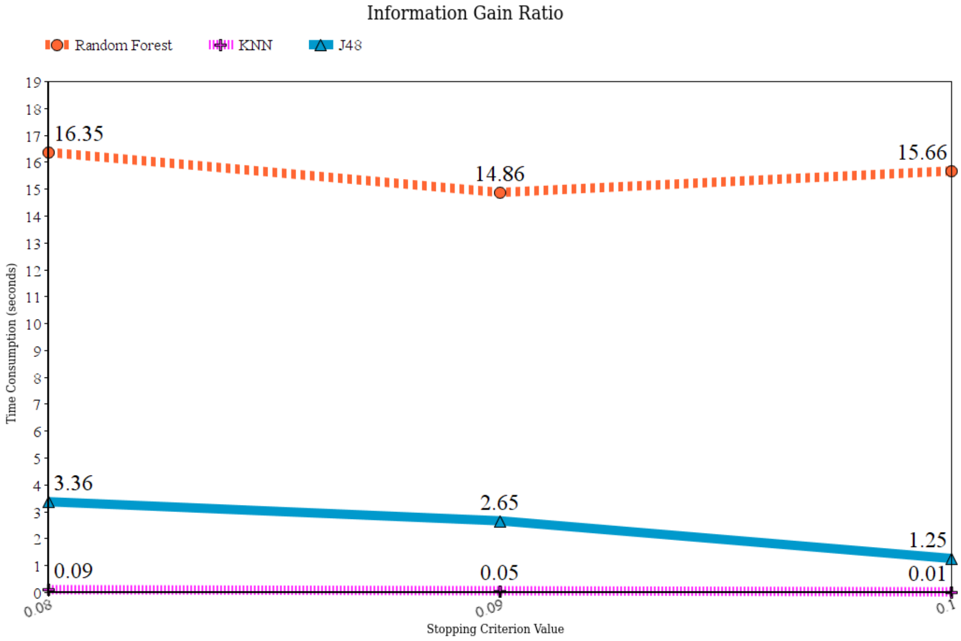

FIGURE 4. Execution time using information gain

most features (the smallest value of stopping criterion). The best accuracy percentage in information gain attribute evaluation is 99.932% (the k-NN method and stopping criterion of 0.08). We know from those figures that the best detection system needs 1 second for processing.

Nevertheless, a slightly different pattern happens to Pearson's correlation that by using 0.2 as the stopping criterion, it is to be the peak and bottom of the accuracy, depending on what classifier being used. The evaluation in Pearson's correlation shows that the best percentages in J-48 and k-NN classification methods are on stopping criteria value of 0.2, but in RF it is 0.1. We find that there are ineffective features that have been selected.

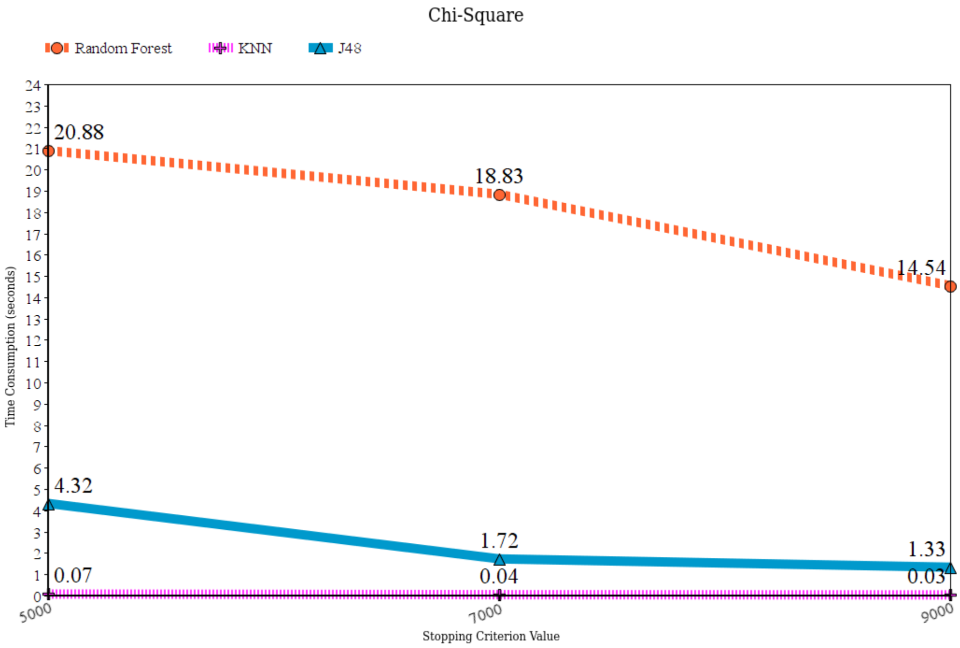

Figure 5. Execution time using Chi-square

Table 3. Accuracy of the proposed method

Dataset	Information gain			Pearson's correlation			Chi-square		
	0.08	0.09	0.1	0.1	0.2	0.3	5000	7000	9000
RF	99.927	99.890	99.885	99.943	99.911	99.927	99.927	99.927	99.927
k-NN	99.932	99.890	99.885	99.917	99.922	99.906	99.937	99.937	99.917
J-48	99.906	99.812	99.823	99.854	99.904	99.864	99.880	99.880	99.880

In this evaluation, we reach the accuracy of 99.943% from the RF classification method, with 0.1 of the stopping criterion. In this detection, we need 25 seconds to process.

In the Chi-square attribute evaluation, the stopping criterion does not much affect the accuracy. It is stable for various classifiers and statistical methods, except the k-NN with 9000 of the stopping value. It can happen with the possibility of features that affect the calculation of learning in the k-NN method. In this detection system, we reach the best percentage of accuracy, 99.937% (the processing time is 0.04 seconds), using the k-NN classifier with the stopping criterion of 7000.

4. **Conclusions.** This paper has proposed feature selection using feature rank and the detection using some machine learning algorithms. We detect the DDoS attacks, especially the PortMap, as a reflection-based DDoS attack. Here, the first is to select 10% of the data. The next step is that we select the features to reduce the dimensional data using feature rank. The last step is classification using a machine learning algorithm.

In this paper, we use random forest, k-nearest neighbor, and J-48 methods. Our proposed method reaches the most effective result in 99.937% of accuracy, which takes 0.04 seconds from the Chi-square attribute evaluation with stopping criteria of 7000 and k-NN classifier.

In the future, research can be conducted to improve current performance and its usability. It includes other classifiers, which may be more effective to implement according to the respective environment. Moreover, feature selection should also consider other types of DDoS attacks to make the method more applicable.

Acknowledgment. The authors gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work under the Publication Writing and IPR Incentive Program (PPHKI) scheme.

REFERENCES

[1] R. K. Deka, D. K. Bhattacharyya and J. K. Kalita, Active learning to detect DDoS attack using ranked features, *Comput. Commun.*, vol.145, pp.203-222, doi: 10.1016/j.comcom.2019.06.010, 2019.

[2] M. Alharbi and M. A. Albahar, Time and frequency components analysis of network traffic data using continuous wavelet transform to detect anomalies, *International Journal of Innovative Computing, Information and Control*, vol.15, no.4, pp.1323-1336, doi: 10.24507/ijicic.15.04.1323, 2019.

[3] A. Dahiya and B. B. Gupta, Multi attribute auction based incentivized solution against DDoS attacks, *Comput. Secur.*, vol.92, 101763, doi: 10.1016/j.cose.2020.101763, 2020.

[4] I. Sharafaldin, A. H. Lashkari, S. Hakak and A. A. Ghorbani, Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy, *Proc. of Int. Carnahan Conf. Secur. Technol.*, doi: 10.1109/CCST.2019.8888419, 2019.

[5] B. Brenner, NetBIOS, RPC PortMap and sentinel reflection DDoS attacks, *Akamai.com*, https://blogs.akamai.com/2015/10/netbios-rpc-portmap-and-sentinel-reflection-ddos-attacks.html, Accessed on August 27, 2020.

[6] M. Aamir and S. M. A. Zaidi, Clustering based semi-supervised machine learning for DDoS attack classification, *J. King Saud Univ. – Comput. Inf. Sci.*, vol.33, no.4, pp.436-446, doi: 10.1016/j.jksuci.2019.02.003, 2021.

[7] S. La and N.-W. Cho, A study on evaluation measures for unsupervised outlier detection, *ICIC Express Letters*, vol.14, no.5, pp.515-520, doi: 10.24507/icicel.14.05.515, 2020.

[8] F. A. Mazarbhuiya, M. Y. Alzahrani and A. K. Mahanta, Detecting anomaly using partitioning clustering with merging, *ICIC Express Letters*, vol.14, no.10, pp.951-960, doi: 10.24507/icicel. 14.10.951, 2020.

[9] S. M. Kasongo and Y. Sun, A deep learning method with wrapper based feature extraction for wireless intrusion detection system, *Comput. Secur.*, vol.92, 101752, doi: 10.1016/j.cose.2020.101752, 2020.

[10] M. M. Hassan, A. Gumaei, A. Alsanad, M. Alrubaian and G. Fortino, A hybrid deep learning model for efficient intrusion detection in big data environment, *Inf. Sci. (Ny.)*, vol.513, pp.386-396, doi: 10.1016/j.ins.2019.10.069, 2020.

[11] Canadian Institute for Cybersecurity, *DDoS 2019 Datasets Research Canadian Institute for Cybersecurity UNB*, 2019, https://www.unb.ca/cic/datasets/ddos-2019.html, Accessed on August 04, 2020.

[12] M. N. Aziz and T. Ahmad, Cluster analysis-based approach features selection on machine learning for detecting intrusion, *Int. J. Intell. Eng. Syst.*, vol.12, no.4, pp.233-243, doi: 10.22266/ijies2019.0831.22, 2019.

[13] Y. Sugianela and T. Ahmad, Pearson correlation attribute evaluation-based feature selection for intrusion detection system, *Proc. of Int. Conf. Smart Technol. Appl.*, Surabaya, Indonesia, pp.1-5, doi: 10.1109/ICoSTA48221.2020.1570613717, 2020.

[14] C. A. Kumar, M. P. Sooraj and S. Ramakrishnan, A comparative performance evaluation of supervised feature selection algorithms on microarray datasets, *Procedia Comput. Sci.*, vol.115, pp.209-217, 2017.

[15] S. Tandon, Entropy: How decision trees make decisions | by Sam T | towards data science, *Towards Data Science*, 2019, https://towardsdatascience.com/entropy-how-decision-trees-make-decisions-2946b9c18c8, Accessed on September 12, 2020.

[16] I. S. Thaseen and C. A. Kumar, Intrusion detection model using fusion of Chi-square feature selection and multi class SVM, *J. King Saud Univ. – Comput. Inf. Sci.*, vol.29, no.4, pp.462-472, doi: 10.1016/j.jksuci.2015.12.004, 2017.

[17] Canadian Institute for Cybersecurity, *NSL-KDD Dataset*, 2009, https://www.unb.ca/cic/datasets/nsl.html, Accessed on September 11, 2020.