Angga/Correlation-Based_Feature_Selection_on_Botnet_Activity_Detection_Using_Kendall_Correlation.pdf

By Angga Pradipta

WORD COUNT	4919	TIME SUBMITTED	25-SEP-2023 10:39AM
		PAPER ID	102984554

Correlation-Based Feature Selection on Botnet Activity Detection Using Kendall Correlation

Dandy Pramana Hostiadi
Department of Informatics
Institut Teknologi dan Bisnis STIKOM Bali
Denpasar, Indonesia
dandy@stikom-bali.ac.id

Yohanes Priyo Atmojo
Department of Informatics
Institut Teknologi dan Bisnis STIKOM Bali
Denpasar, Indonesia
yohanes@stikom-bali.ac.id

Roy Rudolf Huizen
Department of Informatics
Institut Teknologi dan Bisnis STIKOM Bali
Denpasar, Indonesia
roy@stikom-bali.ac.id

I Made Darma Susila
Department of Informatics
Institut Teknologi dan Bisnis STIKOM Bali
Denpasar, Indonesia
darma_s@stikom-bali.ac.id

Gede Angga Pradipta
Department of Informatics
Institut Teknologi dan Bisnis STIKOM Bali
Denpasar, Indonesia
angga_pradipta@stikom-bali.ac.id

I Made Liandana
Department of Informatics
Institut Teknologi dan Bisnis STIKOM Bali
Denpasar, Indonesia
liandana@stikom-bali.ac.id

***Abstract*— Botnets are a dangerous threat to computer networks that uses malicious code to infect computer networks. Thus, the right system security model is needed to detect botnet attack activities accurately. Several previous studies have introduced a botnet detection model using mining-based, but it requires the correct approach to obtain the optimal performance. This paper proposes a botnet detection model by improving feature selection using correlation-based analysis. The aim is to improve accuracy detection by analyzing features with solid correlations that can be used for machine learning classification models. The proposed model consists of 4 main parts: data splitting pre-processing, classification process, and evaluation. The experiment used public datasets, namely CTU-13 dataset containing botnet activity. The experiment shows that the model can detect botnet activity with a detection accuracy of 99.7218%, precision of 99.1691%, and recall of 96.6533%. The proposed model can improve the existing botnet detection system model.**

Keywords—Botnet, Bot activity, Bot Detection, Network Security.

I. INTRODUCTION

System security requires serious attention in the cyber era [1]. An intrusion Detection System (IDS) is known as a security system that is currently widely used as an attack handling [2], [3]. Along with the development of technology, attacks develop into dangerous forms of activity such as involving illegal software called malware [4], [5].

The threat of malware in the cyber era is overgrowing [6]. Malware tends to use computers that have been infected to carry out malicious activities, which are called botnets [7], [8]. A botnet consists of a collection of computers that have been infected and form a communication network. The botnet consists of a master bot and a client bot [9]. The master bot controls each bot client to attack the target computer. Some of the dangerous activities of botnets include Spam activity, click ad fraud, identity theft, spreading malicious code programming, Denial of Service attacks (DDoS), phishing, and adware illegal installation [10], [11].

In its development, botnets have several characteristic behavioral patterns, such as centralized, distributed, and spreading [12]. Thus, it takes the right detection technique to detect bot activity accurately. Several detection techniques that can be used include DNS-based [13], [14], mining-based [6], [10], [15], anomaly-based [8], [16] and signature-based [17], [18]. However, it must be optimized through proper feature selection techniques.

Feature selection is part of the pre-processing stage used in modeling to reduce feature dimensions [2], [11], [14]. In addition, feature selection can be used to increase detection accuracy [2], [19], [20]. In previous studies, botnet detection has resulted in high detection accuracy with feature selection techniques such as the use of Principal Component Analysis (PCA) methods [21], [22], lightweight Logistic Regression model [23] and manuals based on model requirements [9], [11], [15], [24]. However, it has not shown a correlation between each feature and feature priority to improve detection accuracy in the botnet activity detection model. Correlation analysis between each feature is needed to find a strong relationship between one feature and another to improve the accuracy of the detection model performance.

This paper proposes a botnet detection model by optimizing the feature selection process using correlation analysis. The aim is to improve detection accuracy through correlation analysis between features using the Kendal Correlation algorithm. In correlation analysis, a threshold value is used to determine the strong correlation between features. The selected features are used in the *k* -NN classification machine learning model. The *k*-NN method is used because it does not require complex parameters, is easy to implement, and can obtain optimal accuracy detection [25].

The paper is organized into several sections. Previous studies related to feature selection techniques in botnet detection models are described in Section II. Section 3 introduces the process stages of the proposed model. The trials and results of the research are presented in Section IV. Finally, section V presents the conclusions of the research.

II. RELATED WORK

The botnet activity detection model using a mining-based approach is popular to use. Generally, to improve detection performance, optimization of feature selection is carried out. Feature selection techniques have been developed in many botnet detection models [2], [8], [11], [14], [22], [23].

2022 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM) | 978-1-6654-7650-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/CENIM56801.2022.10037525

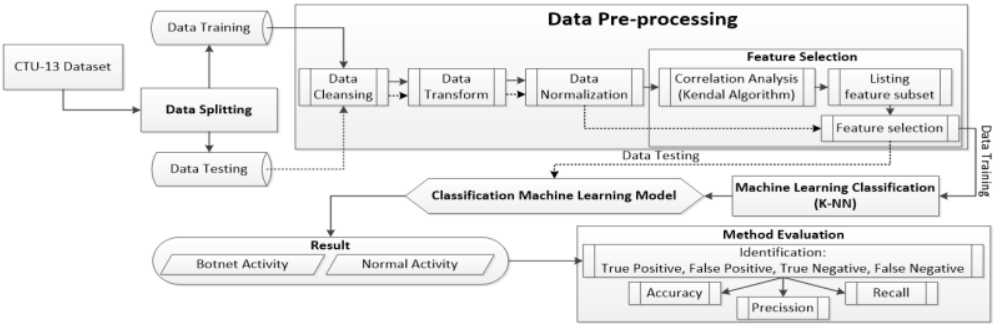

Fig. 1. Proposed Methodolgy

Alieyan et al. [26] proposed a botnet attack detection model based on DNS features. In the feature selection process, two algorithms are used: Information Gain Ratio (IGR) and Principal Component Analysis (PCA) algorithm, and 9 out of 19 features were selected based on the intersection of measurements of the two feature selection methods. The nine selected features are time, source IP address, destination IP address, QNAME, QR, RCODE, domain length, packet length, and TTL (domain response). This study has a high level of rationality to increase detection accuracy in the botnet activity detection model. However, the proposal for using the feature selection method has not been implemented.

Hostiadi, Wibisono and Ahmad [27] developed a botnet detection model by manually selecting 8 out of 13 features in the CTU-13 dataset [28]. The eight features in question are duration, protocol, source port, destination port, source IP, destination IP, total packets, and total bytes. The result of botnet activity detection accuracy obtained is 89.16%. In addition, the detection model named B-corr model can detect attack behavior as a bot group activity. However, the research has not analyzed how strong the relationship between the features represented in the form of a correlation relationship is.

The IoT Botnet detection model was introduced using several feature selection techniques, including mutual information (MI), PCA, and ANOVA F-test in [29]. Of the three feature selection methods, MI is the best feature selection technique capable of producing detection accuracy in the classification process. The highest detection accuracy in the MI feature selection technique is 99.903% in the k-NN classification method. However, the feature selection results have not shown a strong correlation between features.

Velasco-Mata et al. [25] introduced a botnet activity detection model using two filter techniques for feature selection. The two techniques in question are Gini Importance (GI) and Information Gain (IG). There are five best features selected from the feature selection stage. The results of the two feature selection methods produce model performance through an F1 score of 94%, with the best classification method being the Decision Tree. The results of the F1 score are highly but have not analyzed the relationship between features that strongly affect the accuracy of botnet activity detection.

III. Methodology

This paper proposes a feature selection technique using Kendal correlation in the botnet detection model. The purpose of correlation analysis is to obtain an analysis of the effect of correlated feature pairs on the performance of the botnet detection model. In this paper, the research methodology is shown in Fig. 1.

A. Problem Definition and Notation

The feature number and dimensions on a dataset can be reduced by feature selection [30]–[32]. The feature selection algorithm's advantage is that it can improve the detection performance of attack detection models in computer networks [33]–[35]. Besides, the use of appropriate feature selection techniques can reduce computational time [31], [32], [36], [37]. Based on this, this paper analyzes correlation-based feature selection techniques using the Kendall correlation algorithm. The aim is to see the effect of the detection performance of the detection model, measured from detection accuracy, recall, and precision.

The feature selection process in this paper adopts correlation analysis using the Kendall Rank Correlation Coefficient. Kendall correlation is one of the algorithms for measuring the correlation between two sets of ratings given to the same set of objects [38]. The strength of similarity as a correlation between two feature sets is measured and ranked to obtain correlated features. Furthermore, the correlation results are used in the classification process using k-NN. In this paper, notation is used to describe the proposed model.

- Record ($\mathcal{G}$). The dataset has a collection of data records (φ). Thus it is written as $\varphi \in \mathcal{G}$, $\mathcal{G} = \{\varphi_0, \varphi_1, \dots, \varphi_j\}$.
- Feature (Φ). Φ is consists of feature sets (ϕ). If feature (ϕ) is an element of φ, denoted as Φ, written as $\Phi = \{\phi_0, \phi_1, \dots, \phi_j\}$, then $\Phi \in \varphi$, $\varphi = \{\phi_0, \phi_1, \dots, \phi_j\}$.
- Correlation (τ). The correlation between f_i and f_j is calculated using the Kendall correlation equation (1).

$$\tau = \frac{2\,(C-D)}{\sqrt{n(n-1)-T_x}\sqrt{n(n-1)-T_y}} \quad (1)$$

where τ is the correlation coefficient, C is the number of pairs that are in the same direction, D is the number of pairs that are in the opposite direction, n is the number of pairs of x and y, T_x is the ranking correction factor for x and T_y is the ranking correction factor for y.

B. CTU-13 Dataset

The Czech Technical University owns CTU Public dataset through a lab project called the Stratosphere IPS Laboratory.

This dataset contains network traffic at the Czech Technical University containing malware activity. The CTU and malware capture dataset consist of a pure dataset of Botnet malware activity, normal traffic, and a combination of network traffic contaminated with Botnet malware activity or normal activity on the Czech Technical University network. Several different botnets were built and recorded at CTU University in 2011 and are known as the CTU-13 Dataset [28].

C. Data Splitting

At this stage, the dataset is divided into two parts: training data and testing data. We use 70% as training data and 30% as testing data. Then the two data are continued in the data pre-processing process.

D. Data Pre-processing

The pre-processing stage consists of four stages, beginning with the data cleansing process. The data cleansing process standardizes each feature's values and deletes data records. Some values in the feature do not have standardized writing, such as a writer from the SrcAddr feature in IPv6 form, thus requiring a written change to IPv4. In addition, there is an empty feature value (null), and the data record is deleted.

The second stage is data transformation, changing categorical data into numeric data. In this paper, the data transformation used one hot encoding, the same technique as in [12]. The third process is the data normalization process. Each feature value is numeric data that has various value ranges. So normalization data is needed to uniform the range of values in each feature. In this study, normalization was carried out using a value pool of 0 to 1, where 0 was the lowest value limit and 1 was the highest value limit. Then after normalizing the data, the fourth process is carried out, namely the selection feature on the training data. In contrast, the testing data is prepared as a classification process after the machine learning model is formed.

The feature selection begins with determining the number of feature pairs. The featured pair is formed by using the combination in (2).

$$comb = \frac{\phi!}{(\phi-e)!e!}, \tag{2}$$

where $comb$ is the number of features, which are 14 features, and e is the number of selected features, which are eight features. Then the four features are calculated using the Kendall correlation by adopting (1), so that it becomes (3).

$$\tau(\phi_i, \phi_j) = \frac{2(C-D)}{\sqrt{n(n-1)-T_{\phi i}}\sqrt{n(n-1)-T_{\phi j}}} \tag{3}$$

where $\tau(\phi_i, \phi_j)$ is the correlation coefficient between the first feature ϕ_i and the second ϕ_j, C is the number of pairs of features ϕ that are in the same direction, and D is the number of pairs of features ϕ in the opposite direction. n is the number of pairs of ϕ_i and ϕ_j, $T_{\phi i}$ is the ranking correction factor ϕ_i and $T_{\phi j}$ is the ranking correction factor ϕ_j.

The correlation strength between ϕ_i and ϕ_j is determined using the correlation threshold value (4).

$$threshold_\tau = \frac{min_\tau + max_\tau}{2}, \tag{4}$$

threshold τ is the correlation threshold between ϕ_i and ϕ_j, min_τ is the minimum correlation value, and max_τ is the maximum correlation value. Each feature with a strong correlation is based on the correlation threshold value, followed by an analysis of its occurrence in each feature pair set. The result is a sequence of features that appear the most and become the selected features for the machine learning model.

E. Machine Learning Classification Model

In the pre-processing stage, two data are used: training data and testing data at the step of forming a classification model using training data. The formation of the machine learning model in this paper uses the k-NN classification method. The value of k uses 5, calculated using (5).

$$\delta(a,b) = \sqrt{\sum_{k=1}^{d}(X_k - Y_k)^2}, \tag{5}$$

where $\delta(a,b)$ is the proximity between two feature vectors, d is the number of vectors, k is the length of the vector, X is the first vector data, and Y is the second vector data. After the machine learning classification model is formed, then the classification of the test data is carried out. The result of the classification is the detection of attack activity contained in the dataset. Then the performance of classification results is evaluated in the evaluation process.

F. Performane Evaluation

Measurement of model detection performance uses a confusion matrix, where true positive (TP), false positive (FP), false negative (FN), and true negative (TN) values are traced from the detection results of the k-NN machine learning model. TP is the number of botnet activities detected as a botnet attack. FP is the amount of normal activity data detected as a botnet attack. FN is the amount of data on botnet activity that is not detected, and TN is the number of normal activities detected as normal. The TP, FP, FN, and TN values are summarized in the confusion matrix table in Table 1.

TABLE I. CONFUSION MATRIX EVALUATION

Predicted Value \ Actual	True	False
True	TP	FP
False	FN	TN

From the search for confusion matrix values, accuracy, precision, and recall are calculated in equations (6), (7), and (8).

$$Accuracy = \frac{TP + TN}{TP+FP+FN+TN} \tag{6}$$

$$Precision = \frac{TP}{TP+FP} \tag{7}$$

$$Recall = \frac{TP}{TP+FN} \tag{8}$$

IV. EXPERIMENT AND RESULT

In this study, it has a hardware environment with a core i5-7200U processor, 8 GB RAM, and 500 GB storage. The model is built with three programming languages supported by several libraries such as NumPy, pandas, and sci-kit-learn. The test in this study uses the CTU-13 dataset by selecting scenario 9. Description of the dataset is shown in Table 2.

The data cleansing process reduced the number of records by 2,4103%, or as many as 66,376 records in the normal

TABLE II. NSL-KDD DATASET DESCRIPTION

Dataset	Records	Normal	Botnet	Number of Features
Data Description (CTU-13 Scenario 9)	2,753,884	2,574,004	179,880	14
Data Training (70%)	1,927,719	1,801,803	125,916	
Data Testing (30%)	826,165	772,201	53,964	

TABLE III. FEATURE FREQUENCY

Parameter	Rank Feature										
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
Feature	TotPkts	SrcBytes	TotBytes	Dport	State	DstAddr	SrcAddr	sTos	dTos	Proto	Sport
Frequency	27	22	21	16	12	11	6	6	6	6	4

TABLE IV. PRE-PROCESSING RESULT

Dataset (CTU-13 Scenario 9)	Before Pre-processing				After Pre-processing				Reduce percentage (%)			
	Records	*Normal*	*Attack*	*Number of Feature*	*Records*	*Normal*	*Attack*	*Number of Feature*	*Records*	*Normal*	*Attack*	*Features*
Data Description	2,753,884	2,574,004	179,880	14	2,687,508	2,507,628	179,880	11	2.4103	2.5787	0	21.4286
Data Training (70%)	1,927,719	1,801,803	125,916		1,881,256	1,755,340	125,916		2.4103	2.5787	0	
Data Testing (30%)	826,165	772,201	53,964		806,252	752,288	53,964		2.4102	2.5787	0	

activity class label. This reduction affects the composition of the amount of training data and the amount of testing data. In addition, one feature is removed or ignored at the data cleansing stage, namely the Starttime feature. The reason is that this model does not take into account the analysis of activity time. Then proceed to the data transformation process. Three features are changed from categorical data to numerical data, namely Proto, Dir, and State.

The data transformation results from the value in each feature into a numerical basis with various value ranges. In this study, the normalized data changed the range of values in each feature on a scale of 0 to 1. The 0 value indicates the smallest value range, and 1 is the highest range of values. Then the results of the normalized data feature selection.

The correlation measurements using Kendall Correlation obtained the lowest correlation value of 0.0018, which is taken from features Dir, State, Dur, and dTos. Besides, the highest value of 0.3029 is taken from TotPkts, SrcBytes, TotBytes, and Dport. Thus, to determine the strong correlation between feature sets is 0.1523. Based on the correlation threshold, there are 104 combinations of feature sets, with the number of each set being eight features. Then an analysis of the occurrence of the same feature is carried out in pairs of different feature sets. As a result, 11 features strongly correlate, namely the TotPkts, SrcBytes, TotBytes, Dport, State, DstAddr, SrcAddr, sTos, dTos, Proto, and Sport features. The results of the feature occurrence analysis are shown in Table 3.

Feature selection is the last stage in pre-processing. The pre-processing results can reduce the data records and the number of features. Details of the percentage reduction in pre-processing are shown in Table 4.

The eleven features selected in the feature selection process are used in the classification process using the k-NN model. The classification results produce a confusion matrix value shown in Table 5.

TABLE V. CONFUSION MATRX VALUE

Predicted Value	Actual value: True	Actual value: False
True	52,158	437
False	1,806	751,851

The detection results show that the classification model can detect botnet activity with a detection accuracy of 99.7218%, precision of 99.1691%, and recall of 96.6533%. These results indicate that the detection model has an accurate detection performance influenced by the feature selection process. The features used in the classification process use *k*-NN machine learning using 11, which has a strong correlation and has a high frequency of occurrence in feature combinations. The measurement results are shown in Table 6.

TABLE VI. MODEL EVALUATION

TP	FP	FN	TN	Acc.	Prec.	Rec.
96.6533	0.0581	3.3467	99.9419	99.7218	99.1691	96.6533

In addition to testing the detection performance, in this paper, processing analysis is carried out to see the time it takes to get the detection results. The results of the time measurement are shown in Fig. 2.

The computational time measurement results show a high processing time consumption in the feature selection and classification process. This is because, at the time of selection, the process of forming a combination of feature sets and measuring correlations is carried out. Besides, the classification process requires processing time to classify testing data using the *k*-NN classification model. In this paper, the proposed model's detection results are compared with those in previous studies. Table 7 shows that the model has

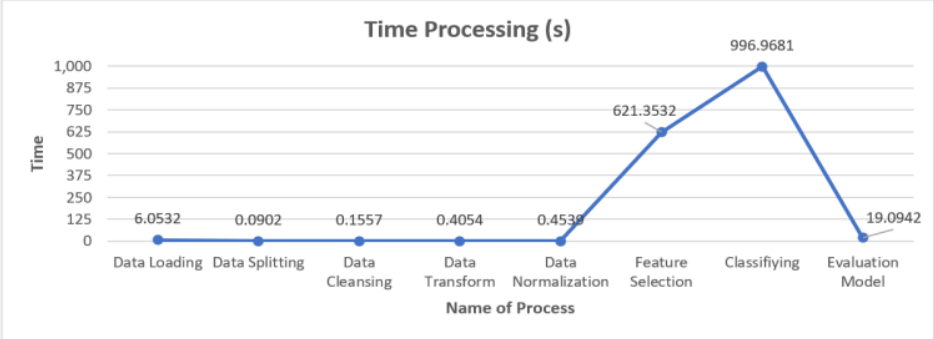

Fig. 2. Computation Time

TABLE VII. COMPARATION RESULT

Model	Acc.	Prec.	Rec.	Dataset	Correlation analysis	Time Analysis
Hostiadi and Ahmad [1]	99.18%	42.29%	91.55%	CTU-Dataset	√	-
Dollah et al. [39] • Decission Tree • *k*-NN • Naïve Bayes • Random Forest	 92.20 75.16 69.34 73.83	 99.93 73.18 62.28 49.99	 84.47 51.52 99.45 47.67	CTU-Dataset	-	-
Eslah, Abidin and Naseri [40] • C4.5 • Random Forest • Naïve Bayes • Support Vector Machine • Feedfprward Neural Network (F-NN)	 98.20 98.20 97.00 98.40 98.50	 98.20 98.20 97.00 98.50 98.50	 98.20 98.20 97.00 98.50 98.50	CTU-Dataset	-	√
Proposed Model	**99.7218**	**99.1691**	**96.6533**	**CTU-Dataset**	**√**	**√**

higher accuracy than research [1],[39],[40]. In terms of precision, the proposed model has a higher value than the research in [1],[39],[40]. But has a lower classification model than the Decision Tree [39]. Recall measurement has the lowest value in the study. In contrast to previous studies, in this paper, the proposed model has the advantage of performing feature selection based on correlation analysis using the Kendall correlation and measuring the computational time that has never been done in previous studies. So, the proposed model can be used to develop a special system security model to detect botnet attacks.

V. CONCLUSION

This paper proposes a new approach to feature selection using correlation analysis to increase detection accuracy in the botnet attack detection model. The proposed model consists of 4 main processes: data splitting, pre-processing, classification and evaluation. The feature selection process successfully reduced the feature dimensional from 11 out of 14 features in pre-processing stages, which have a strong correlation. Besides, it has a high frequency of occurrence in feature set pairs and affects the detection accuracy results. The experiment used the threshold values to determine the set feature with values of 0.1523 to indicate the strong and weak correlation between each selected feature. The detection results show that the classification model has the highest accuracy compared to previous studies, 99.7218%. Compared to previous research, the model has a novelty regarding feature analysis which has a strong correlation with each other and has computational time analysis.

In the future, the proposed model can be developed by examining the use of the classification model. The classification model can maximize precision and recall performance by following the feature selection method.

ACKNOWLEDGMENT

Appreciation is given to the STIKOM Bali Institute of Technology and Business for the support provided so that this research can be completed and improve the quality of our performance.

REFERENCES

[1] D. P. Hostiadi and T. Ahmad, "Hybrid model for bot group activity detection using similarity and correlation approaches based on network traffic flows analysis," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 7, pp. 4219–4232, 2022.

[2] R. Zhao, "A Hybrid Intrusion Detection System Based on Feature Selection and Weighted Stacking Classifier," IEEE Access, vol. 10, no. June, pp. 71414–71426, 2022.

[3] V. Varanasi and S. Razia, "Network Intrusion Detection using Machine Learning, Deep Learning - A Review," Adv. Sci. Technol. Res. J., vol. 16, no. 3, pp. 193–206, 2022.

[4] D. Geer, "Malicious bots threaten network security," Computer (Long. Beach. Calif)., vol. 38, no. 1, pp. 18–20, 2005.

[5] H. Zeidanloo, F. Tabatabaei, P. Vahdani Amoli, and A. Tajpour, "All About Malwares (Malicious Codes)," Secur. Manag. 2010all a, no. January, pp. 342–348, 2010.

[6] X. Dong, J. Hu, and Y. Cui, "Overview of botnet detection based on machine learning," Proc. - 2018 3rd Int. Conf. Mech. Control Comput. Eng. ICMCCE 2018, pp. 476–479, 2018.

[7] A. A. Selcuk, F. Orhan, and B. Batur, "Undecidable problems in malware analysis," 2017 12th Int. Conf. Internet Technol. Secur. Trans. ICITST 2017, pp. 494–497, 2018.

[8] J. Maestre Vidal, A. L. Sandoval Orozco, and L. J. García Villalba, "Alert correlation framework for malware detection by anomaly-based packet payload analysis," J. Netw. Comput. Appl., vol. 97, no. January 2016, pp. 11–22, 2017.

[9] H. Dhayal and J. Kumar, "Botnet and P2P Botnet Detection Strategies: A Review," Proc. 2018 IEEE Int. Conf. Commun. Signal Process. ICCSP 2018, pp. 1077–1082, 2018.

[10] D. P. Hostiadi and T. Ahmad, "Sliding Time Analysis in Traffic Segmentation for Botnet Activity Detection," 5th Int. Conf. Comput. Informatics, ICCI 2022, pp. 286–291, 2022.

[11] M. Hajizadeh, M. A. Jahromi, and T. Bauschert, "An Unsupervised Ensemble Learning Approach for Novelty-based Botnet Detectors," pp. 713–714, 2022.

[12] M. A. R. Putra, T. Ahmad, and D. P. Hostiadi, "Analysis of Botnet Attack Communication Pattern Behavior on Computer Networks," Int. J. Intell. Eng. Syst., vol. 15, no. 4, 2022.

[13] U. Bayer, A. Moser, C. Kruegel, and E. Kirda, "Dynamic analysis of malicious code," pp. 67–77, 2006.

[14] S. Dongre, "Analysis of Feature selection techniques for Denial of Service (DoS) attacks .," 2018 4th Int. Conf. Recent Adv. Inf. Technol., vol. 48, pp. 1–4, 2018.

[15] Z. Chu, Y. Han, and K. Zhao, "Botnet Vulnerability Intelligence Clustering Classification Mining and Countermeasure Algorithm Based on Machine Learning," IEEE Access, vol. 7, pp. 182309–182319, 2019.

[16] M. Prajapati and D. Dave, "Host-based forensic artefacts of botnet infection," Proc. - Int. Carnahan Conf. Secur. Technol., vol. 2019-Octob, pp. 0–3, 2019.

[17] W. Sun and H. Gou, "The botnet defense and control," Proc. - 2011 Int. Conf. Inf. Technol. Comput. Eng. Manag. Sci. ICM 2011, vol. 4, pp. 339–342, 2011.

[18] W. Zhang and J. Lu, "SEIR-based botnet propagation model," Proc. - 2021 6th Int. Conf. Smart Grid Electr. Autom. ICSGEA 2021, vol. 6, pp. 439–442, 2021.

[19] A. A. Megantara and T. Ahmad, "ANOVA-SVM for Selecting Subset Features in Encrypted Internet Traffic Classification," Int. J. Intell. Eng. Syst., vol. 14, no. 2, pp. 536–546, 2021.

[20] E. Pune, E. Pune, and E. Pune, "Ensemble Based Feature Selection Technique For Flow Based Intrusion Detection System .," pp. 3–6, 2022.

[21] K. Alieyan, M. Anbar, A. Almomani, R. Abdullah, and M. Alauthman, "DNS Features," 2018 Int. Arab Conf. Inf. Technol., pp. 1–4, 2018.

[22] A. Muhammad, M. Asad, and A. R. Javed, "Robust Early Stage Botnet Detection using Machine Learning," 1st Annu. Int. Conf. Cyber Warf. Secur. ICCWS 2020 - Proc., 2020.

[23] R. Bapat et al., "Identifying malicious botnet traffic using logistic regression," 2018 Syst. Inf. Eng. Des. Symp. SIEDS 2018, pp. 266–271, 2018.

[24] J. Lu, F. Lv, Q. H. Liu, M. Zhang, and X. Zhang, "Botnet Detection based on Fuzzy Association Rules," Proc. - Int. Conf. Pattern Recognit., vol. 2018-Augus, pp. 578–584, 2018.

[25] J. Velasco-Mata, V. Gonzalez-Castro, E. F. Fernandez, and E. Alegre, "Efficient Detection of Botnet Traffic by Features Selection and Decision Trees," IEEE Access, vol. 9, pp. 120567–120579, 2021.

[26] K. Alieyan, A. Almomani, A. Manasrah, and M. M. Kadhum, "A survey of botnet detection based on DNS," Neural Comput. Appl., vol. 28, no. 7, pp. 1541–1558, 2017.

[27] D. P. Hostiadi, W. Wibisono, and T. Ahmad, "B-Corr Model for Bot Group Activity Detection Based on Network Flows Traffic Analysis," KSII Trans. Internet Inf. Syst., vol. 14, no. 10, pp. 4176–4197, 2020.

[28] S. Garcia, "Modelling the Network Behaviour of Malware To Block Malicious Patterns . The Stratosphere Project : a Behavioural Ips," Virus Bull., no. September, pp. 1–8, 2015.

[29] M. Al-Sarem, F. Saeed, E. H. Alkhammash, and N. S. Alghamdi, "An aggregated mutual information based feature selection with machine learning methods for enhancing iot botnet attack detection," Sensors, vol. 22, no. 1, 2022.

[30] J. Lee, D. Park, and C. Lee, "Feature selection algorithm for intrusions detection system using sequential forward search and random forest classifier," KSII Trans. Internet Inf. Syst., vol. 11, no. 10, pp. 5132–5148, 2017.

[31] O. Joseph, "Network Intrusion Detection Models based on Naives Bayes and C4 . 5 Algorithms," pp. 3–7, 2022.

[32] M. N. Aziz and T. Ahmad, "Clustering under-sampling data for improving the performance of intrusion detection system," J. Eng. Sci. Technol., vol. 16, no. 2, pp. 1342–1355, 2021.

[33] M. N. Reza, S. F. Kabir, N. Jahan, and M. Islam, "Evaluation of Machine Learning Algorithms using Feature Selection Methods for Network Intrusion Detection Systems," 2021 5th Int. Conf. Electr. Inf. Commun. Technol. EICT 2021, no. December, pp. 17–19, 2021.

[34] F. H. Almasoudy, W. L. Al-Yaseen, and A. K. Idrees, "Differential Evolution Wrapper Feature Selection for Intrusion Detection System," Procedia Comput. Sci., vol. 167, no. 2019, pp. 1230–1239, 2020.

[35] N. T. Pham, E. Foo, S. Suriadi, H. Jeffrey, and H. F. M. Lahza, "Improving performance of intrusion detection system using ensemble methods and feature selection," ACM Int. Conf. Proceeding Ser., 2018.

[36] S. Saha, A. T. Priyoti, A. Sharma, and A. Haque, "Towards an Optimal Feature Selection Method for AI-Based DDoS Detection System," pp. 425–428, 2022.

[37] L. Zhang and C. Xu, "A Intrusion Detection Model Based on Convolutional Neural Network and Feature Selection," pp. 162–167, 2022.

[38] H. Abdi, "Kendall Rank Correlation Coefficient," Concise Encycl. Stat., pp. 278–281, 2008.

[39] R. F. M. Dollah, M. A. Faizal, F. Arif, M. Z. Mas'ud, and L. K. Xin, "Machine learning for HTTP botnet detection using classifier algorithms," J. Telecommun. Electron. Comput. Eng., vol. 10, no. 1–7, pp. 27–30, 2018.

[40] M. Eslahi, W. Z. Abidin, and M. V. Naseri, "Correlation-based HTTP Botnet detection using network communication histogram analysis," 2017 IEEE Conf. Appl. Inf. Netw. Secur. AINS 2017, vol. 2018-Janua, pp. 7–12, 2017.

Angga/Correlation-Based_Feature_Selection_on_Botnet_Activity_Detection_Us...

ORIGINALITY REPORT

7%
SIMILARITY INDEX

PRIMARY SOURCES

1	www.researchgate.net Internet	83 words — 2%
2	Yohanes Priyo Atmojo, I Made Darma Susila, Muhammad Riza Hilmi, Erma Sulistyo Rini, Lilis Yuningsih, Dandy Pramana Hostiadi. "A New Approach for Spear phishing Detection", 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT), 2021 Crossref	80 words — 2%
3	Dandy Pramana Hostiadi, Made Darma Susila, Roy Rudolf Huizen. "A New Alert Correlation Model Based On Similarity Approach", 2019 1st International Conference on Cybernetics and Intelligent System (ICORIS), 2019 Crossref	67 words — 2%
4	Dandy Pramana Hostiadi, Tohari Ahmad. "Sliding Time Analysis in Traffic Segmentation for Botnet Activity Detection", 2022 5th International Conference on Computing and Informatics (ICCI), 2022 Crossref	59 words — 2%

EXCLUDE QUOTES	ON	EXCLUDE SOURCES	< 2%
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF