Feature selection to improve distributed denial of service detection accuracy using hybrid N-Gram heuristic techniques

Andi Maslan[1], Abdul Hamid[2], Dedy Fitriawan[3], Anggia Dasa Putri[1], Tukino[1]
[1]Department of Informatic Engineering, Faculty of Engineering and Computer Science, Putera Batam University, Batam, Indonesia
[2]Department of Technology Studies, Faculty of Technical and Vocational Education, Universiti Tun Hussein Onn Malaysia, Johor, Malaysia
[3]Department of Remote Sensing and Geographic Information System, School of Vocational, Universitas Negeri Padang, Padang, Indonesia

Article Info

Article history:

Received Dec 24, 2023
Revised May 6, 2024
Accepted May 26, 2024

Keywords:

Chi-square distance
Distributed denial of service
Malware
N-Grams
Payload

ABSTRACT

Distributed denial of service (DDoS) attacks servers and computers in various ways, such as flooding traffic. There are three DDoS detection methods, namely anomaly-based, pattern-based and heuristic-based. However, pattern-based methods cannot detect recent attacks, while anomaly-based methods have low accuracy and relatively high false positives. This research proposes increasing accuracy using a heuristic-based DDoS detection method and a new feature. The combination of CSDPayload+N-Gram and CSPayload+N-Gram features is called hybrid N-Gram, which is analysed on four datasets: CIC2017, CIC2019, MIB-2016, and H2NPayload. Next, calculate Chi-square distance (CSD) and cosine similarity (CS) using the N-Gram frequency value results. Subsequently, compute Pearson Chi-square using the N-Gram frequency value results. Compare the CSDPayload+N-Gram and CSPayload+N-Gram, along with the Pearson Chi-square value, to classify it as either DDoS or not. Finally, feature selection based on weight correlation and payload classification employs machine learning algorithms: support vector machine (SVM), K-nearest neighbors (KNN), and neural network (NN). The average accuracy rate for detecting DDoS attacks across four datasets, utilising the CSDPayload+4-Gram and CSPayload+4-Gram features with the SVM algorithm, is 99.71%, which surpasses the accuracy achieved by using KNN (96.22%) and NNs (99.50%) imitation. Thus, the best algorithm for detecting DDoS is SVM with hybrid 4-Gram.

Corresponding Author:

Andi Maslan
Department of Informatics Engineering, Faculty of Engineering and Computer Science
Putera Batam University
Muka Kuning Batam, Kepulauan Riau, Batam, Indonesia
Email: lanmasco@gmail.com

1. INTRODUCTION

Since data protection in a business is now required, network security is a crucial component. It includes corporate secrecy, after all. The accessibility of the data at the time of access is one important factor. However, occasionally, server disruptions‒such as distributed denial of service (DDoS) attacks‒cause the data to become unavailable. Denial-of-service (DoS) attacks use the internet to target vital websites. By delivering undesired traffic to the victim (computer or network), this attack seeks to deteriorate standard services from legitimate services by using bandwidth or connection capacity. The surge in DoS attacks has significantly raised the risk to servers and network devices on the internet.

Furthermore, there are two issues with the pattern identification of DDoS attacks on intrusion detection system's (IDS). Furthermore, there are two issues with the pattern identification of DDoS attacks in IDS. Firstly, a transmission control protocol/internet protocol (TCP/IP) deficiency makes DDoS attacks easy to launch and makes it difficult to identify victims. In addition, several agents launch DDoS attacks on a single target [1]. Furthermore, DDoS attacks have evolved a new tactic; the SYN-Flood attack is one example [2]. A solitary SYN packet is typically a legitimate packet of network activity that is challenging for intrusion detection systems to identify as an odd artefact. As a result, IDS is difficult enough to produce a warning regarding potential network attacks using SYN-Flood. Second, typical network patterns are often mistakenly recognised as DDoS attacks, leading to false-positive alarm difficulties in signature-based intrusion detection systems. Therefore, in the event of a DDoS assault, it is critical to promptly detect and implement mitigation strategies to safeguard networks that are unable to operate as intended.

The sort of resource depletion assault, for instance, suggests payload-based signature generation as an alternative to the similarity-based classification technique, which interprets payloads as strings, in order to detect DDoS attacks based on the similarity of the two payloads [3]. It looks into ways to correlate the payloads according to their similarity in content and structure [4]. Related payloads that are part of an assault but have a different version from other traffic are intended to be grouped by this classification.

A study by Zhao *et al.* [5] investigated the use of N-Gram approaches to distinguish attacks from benign HTTP traffic, with the N-Gram methodology being implicated in resource depletion. In this study, the research findings were compared with those of the hidden Markov model (HMM)-based methodology [6]. These N-Gram techniques underwent extensive testing on publicly available datasets and simulated traffic, including a highly realistic attack dataset. The results indicate that each approach can achieve a comparable detection rate. The pattern-matching technique was highly efficient in terms of per-packet processing time. However, this study focuses on the number and size of packets without analysing the hex payload in depth in a data packet because the study has limitations on header packet research.

Since the first DDoS attacks in 1990 and 2000, websites of major businesses like Amazon, convolutional neural network (CNN), eBay, and Yahoo have suffered significant downtime lasting several hours. These incidents prompted ongoing research in network security to prevent such attacks. Various techniques, including statistical analysis, knowledge-based methods, software computing, data mining, and machine learning, have been developed to identify and mitigate DDoS attacks [7]. Similar to earlier studies, byte-level HTTP traffic analysis provides a workable approach for network intrusion detection and traffic analysis issues.

Research conducted by [4], [8], [9] led to the development of an intelligent system for detecting DDoS attack patterns using network packet analysis and machine learning techniques. This study analysed numerous network packets provided by the Center for Applied Internet Data Analysis. The researchers implemented a detection system utilising the support vector machine (SVM) algorithm, primarily focusing on the Radial Kernel (Gaussian) function. This study prepared 4,000 IP addresses consisting of 2,000 IP addresses from the attacker pool and 2,000 from the victim pool as test data and four features. The detection system can detect DDoS attacks with an accuracy rate of 85% with all types of data sets and 98.7% accuracy with five features. The strategy for developing a DDoS attack detection system shows that the system with SVM is trained using the proposed features to successfully detect DDoS attacks with high accuracy.

In addition [10], [11] suggested enhanced DDoS attack detection utilising flow-based analysis and the rapid entropy approach. While retaining acceptable detection accuracy, fast entropy and flow-based computing significantly cut computation time compared to conventional computing. Analysis is done on network traffic, and request entropy per stream is computed dynamically. When the flow count entropy and the average entropy value during that time interval differ by a threshold value–which is adaptively modified based on traffic pattern conditions to increase detection accuracy–a DDoS assault is identified. This paper suggests three techniques for DDoS detection: flow aggregation, adaptive threshold, and fast entropy. The adaptive threshold approach decreases computing time and increases detection accuracy compared to standard entropy. For example, the connection between 192.95.27.190 and 71.126.222.64. The resulting value is substantial, namely 7.46 compared to other connections. However, this proposed method performs forward tracking, meaning previously detected packets cannot be analysed again.

Researchers have presented a machine-learning approach to detect DDoS attacks [12]. They gathered a fresh dataset that included contemporary assault types not included in earlier research. There are five classes and 27 characteristics in the dataset. Because the network simulator (NS2) may be employed with high and reasonably reflected findings, it is used in this work [13], [14]. Data for a number of attacks that target the network and application layers has been recorded. To identify Smurf, user datagram protocol (UDP)-Flood, HTTP-Flood, and SIDDOS attacks, three machine learning techniques—multilayer perceptron (MLP), random forest, and Naïve Bayes—were employed on the obtained datasets. The MLP classifier emerged as the most accurate. The experimental results demonstrate that MLP achieved the highest accuracy rate of 98.63%, surpassing random forest and Naïve Bayes.

Then, Niyaz *et al.* [15] describes a deep learning-based DDoS detection system for TCP, UDP, and internet control message protocol (ICMP)-related multi-vector attacks in a specialised software-defined networking (SDN) environment. The proposed approach achieves an accuracy of 95.65% in identifying distinct DDoS attack classes. Compared to previous works, it achieves a 99.82% accuracy rate in classifying traffic as normal and attack classifications, with extremely few false positives. However, as a recommendation for future research, the NIDS system in this study has not been able to identify attacks on the application layer, particularly when dealing with raw data. Maslan *et al.* [16] proposed an intrusion detection system utilising cosine similarity (CS) to address this limitation. So far, the firewall only checks packets based on IP addresses and ports, and IDS works by spreading incoming packets to computers to decide whether incoming packets are malicious. An example of an IDS application is Snort IDS, an open-source application that uses strings to detect malicious activity. One of the disadvantages of string-matching IDS is that the occurrence of strings in a packet must be precisely the same. The slightest difference can make the attack undetectable, making it difficult to detect the same stream but different patterns. Therefore, an intrusion detection method uses CS to find the similarity of several sequence packets. Then, the search is done to find the similarities between the payload and the existing signature.

According to research Sridharan [17] and a follow-up to [6] states that web applications generate malicious HTTP requests that provide a platform for attacking machi; online apps produce fraudulent HTTP requests, which give attackers a platform to target devices that are susceptible to attacks. The network intrusion detection system needs to detect such malicious traffic based on traffic analysis. According to prior studies, the N-Gram approach can detect HTTP attacks. This work uses the Ad-hoc N-Gram technique, pattern counting technique, and Chi-square distance (CSD) to examine the payload size. The study only looks at payload size and 2-to 3-gram comparisons. However, the results indicate that 2-Gram has an AUC value of 0.98 and an accuracy rate of detection of generic assaults, shellcode attacks, and CLET attack dataset of 98.16%.

Research by Zekri *et al.* [18] proposed a signature generation algorithm for detecting DDoS attacks. This algorithm generates a signature based on the attack packets, which are then stored in a database. To distinguish between regular packets and DDoS attacks on the network, this study utilises a string similarity metric to measure the similarity, dissimilarity, or distance between two objects associated with an attack. Each similarity metric is adjusted to produce a number between 0 and 1, where one indicates that the comparison object is identical and 0 indicates that the comparison object is disconnected. In some metrics, this can be done by normalising the metric by dividing the old result by the maximum result. At the same time, for those that resemble distances, a subtraction operation is performed to obtain equality. A separate optimum threshold is used for each similarity metric to increase the metric's accuracy. The detection accuracy rate for CS is 65%, Longest common subsequence 65%, Smith-Waterman similarity 100%, Levenshtein similarity 80%, and Jaccard Index 65%. All the signatures created as a result of this study are then included in the IDS's Snort application.

Furthermore, Aldwairi *et al.* [19] suggested a brand-new, content-based, automated signature-generating approach that creates profile anomalies to find and classify novel yet unreported worms. By producing fewer substrings, the natural tokenisation technique used by the suggested system SCAN accelerates the generation process. In order to address the shortcomings of the old stop word approach's signature substring, this study suggests a new stop character technique. Furthermore, SCANS has an enhanced binary detection model specifically intended to identify attacks. A 95% malicious packet detection rate for port 23 was demonstrated in experimental testing using the DARPA IDS dataset, with specificities of 88.4% and 94.6% for ports 21 and 25, respectively. However, the study was limited to port features and did not analyse the port payload feature in depth.

Then Yulianto *et al.* [20], thought of enhancing the AdaBoost-based IDS performance by the application of ensemble feature selection (EFS), principal component analysis (PCA), and synthetic minority oversampling technique (SMOTE). The process of selecting features involves obtaining weighted value data from the ensemble results for feature selection. The AdaBoost classification, utilised for classification during the training phase, is another important part of this technique. The dataset comes from CIC-2017 [21], which comprises 55,173 labelled as normal and 12,550 labelled as DDoS. This study implements SMOTE with an oversampling minority class of 200%, and for feature selection, a threshold value of T=0.9 is used [22]. The number of minority class instances (DDoS) in the training data increased from 29285 to 87855. 25 features were chosen from a total of 72 based on this approach and method, yielding an accuracy rate of 81.83% for AdaBoost+EFS+SMOTE. Additionally, the effect of the feature selection approach on the ability of machine learning models to detect DDoS attacks was assessed in [23]. Additionally, the report claims that defence methods based on signatures are insufficient to counter new threats like DoS attacks. The primary goal of developing a machine learning classifier is to accurately and efficiently detect DDoS attacks. However, how one selects the "relevant" and "minimal" characteristics in the network flow determines how well machine learning models distinguish DDoS attacks. This study uses five supervised ML-based classifier algorithms: SVM, Gaussian Naïve Bayes (GNB), and K-nearest neighbor (KNN), and random forest. The best algorithm is KNN with a 94% accuracy rate for K=15 and random forest with a 96% accuracy rate. Stiawan *et al*. [24], the proposed control flow graph (CFG) feature detects Ransomware attacks and combines it with the N-Gram feature by extracting the opcode feature. The algorithm used to perform the classification uses KNN. The number of datasets observed in this study was 3,000 normal files and 3,000 ransomware files in windows portable executable (PE) format, samples taken from the VX Heaven virus collection database. The results of this study [25] show that the K-NN classification algorithm can detect malware with the highest accuracy of 98.80%.

Based on the problem background and research motivation, this study aims to design an N-Gram technique to detect DDoS attacks and implement a DDoS detection method based on the N-Gram hybrid heuristic technique. Heuristics, or heuristic techniques, are approaches to problem-solving that use practical methods or various shortcuts to come up with solutions that may not be optimal but are sufficient within a limited time frame or deadline. Heuristic methods are intended to be flexible and used for quick decision-making, especially when finding optimal solutions is impossible or impractical and when working with complex data [26]. In other studies, heuristics develop into metaheuristics to find optimal solutions by considering other factors besides the main ones [27], [28]. However, this research focuses on the N-Gram hybrid heuristic technique.

Based on the research's problems, objectives, and motivation, the first research contribution is the existence of a new dataset called H2N-Payload. The second detects DDoS attacks using payloads extracted from the data packet structure. Payload is specified in hexadecimal, ranging from 1-Gram to 6-Gram. It then calculates the frequency of occurrence of N-Grams using statistical models called CSD and CS. This technique makes two main contributions: detecting DDoS attacks using more than 2-Gram and will result in new features called CSDPayload+N-Gram and CSPayload+N-Gram. The final contribution is to improve detection accuracy using the SVM machine learning algorithm.

2. METHOD

This study uses a Heuristic-based N-Gram technique to detect DDoS attacks. The research phase begins by collecting datasets from CIC-2017 [29], [30], MIB-2016 [31], and a new dataset called the H2Npayload dataset. Then, in the final stage, the three datasets were evaluated with new features by determining the accuracy, precision, recall, F-measure, and ROC level in detecting DDoS attacks. Then, the performance levels of each machine-learning algorithm will be compared. The research steps can be seen in Figure 1.

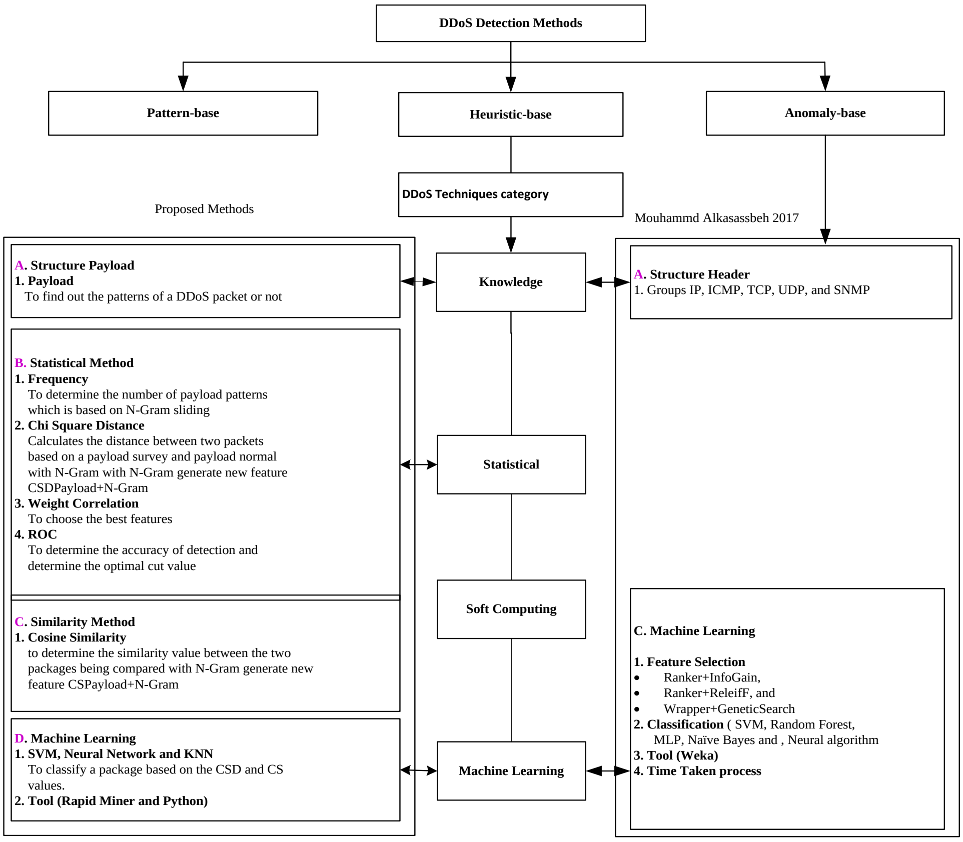

DDoS Detection Methods
Pattern-base
Heuristic-base
Anomaly-base
DDoS Techniques category
Proposed Methods
Mouhammd Alkasassbeh 2017
A. Structure Payload
1. Payload
To find out the patterns of a DDoS packet or not
B. Statistical Method
1. Frequency
To determine the number of payload patterns which is based on N-Gram sliding
2. Chi Square Distance
Calculates the distance between two packets based on a payload survey and payload normal with N-Gram with N-Gram generate new feature CSDPayload+N-Gram
3. Weight Correlation
To choose the best features
4. ROC
To determine the accuracy of detection and determine the optimal cut value
C. Similarity Method
1. Cosine Similarity
to determine the similarity value between the two packages being compared with N-Gram generate new feature CSPayload+N-Gram
D. Machine Learning
1. SVM, Neural Network and KNN
To classify a package based on the CSD and CS values.
2. Tool (Rapid Miner and Python)
Knowledge
Statistical
Soft Computing
Machine Learning
A. Structure Header
1. Groups IP, ICMP, TCP, UDP, and SNMP
C. Machine Learning
1. Feature Selection
Ranker+InfoGain,
Ranker+ReleifF, and
Wrapper+GeneticSearch
2. Classification (SVM, Random Forest, MLP, Naïve Bayes and , Neural algorithm
3. Tool (Weka)
4. Time Taken process

Figure 1. Justification of the research

It can be seen from Figure 1 that the DDoS attack detection method is generally divided into two parts, namely the base pattern and the base anomaly, and can use both, often called the heuristic base method [32]. The heuristic base method is divided into several categories of techniques in detecting DDoS attacks: knowledgebase, statistical base, soft computing base, and machine learning base. Each technique has its algorithm for detecting DDoS attacks. The research can focus on package structures such as headers and payloads if a knowledge base is chosen. If a base statistical model is chosen, various statistical models, such as CSD, correlation, and anova, or parametric and non-parametric statistics, can be utilised.

Soft computing detects DDoS attacks in high-speed applications through complex algorithms and calculations, including fuzzy logic, artificial neural networks, and probabilistic reasoning. Machine learning involves computer programs that learn from experience in specific tasks and performance metrics. Through experiential learning, machine learning programs enhance their performance, making them valuable tools for intelligent decision-making [33].

Heuristic techniques analyse the HTTP protocol and dig deeply into packet data, especially in post commands, get, and other specific commands [34]. Then, the payload will be extracted in hexadecimal form for analysis using the N-Gram technique. The analysis employed two formulas: CSD and CS. CSD calculates the distance between the observed payload and the normal payload, while CS measures the degree of similarity of the observed payload to the normal payload. A value closer to one indicates greater similarity to the comparison payload. This analysis generates new features known as CSDPayload+N-Gram and CSPayload+N-Gram, each assigned a value and threshold to determine whether a packet is malicious.

The network has gathered packet traffic data for a certain amount of time and stored it in PCAP format, which contains distinct data about the length, number, and IP-IP pairs of each packet as well as the payload. A summary of the general attributes can be found in Table 1. Table 1 elucidates that the research utilises two datasets sourced from the internet and one dataset derived from simulated attacks on cloud servers. For experimental purposes, Figure 2 presents an example of normal and subnormal payload results extracted from packets using the scapy module in Python programming.

Table 1. Dataset property

Dataset property	H2NPayload	MIB2016	CICIDS2017	CIC2019
Size dataset	114,5 KB	832 kbps	21 Mpbs	21 Mpbs
Number of features	6	6	6	6
Number of records	1,954	4,998	10,000	10,000
Number of attack type	2	7	2	2

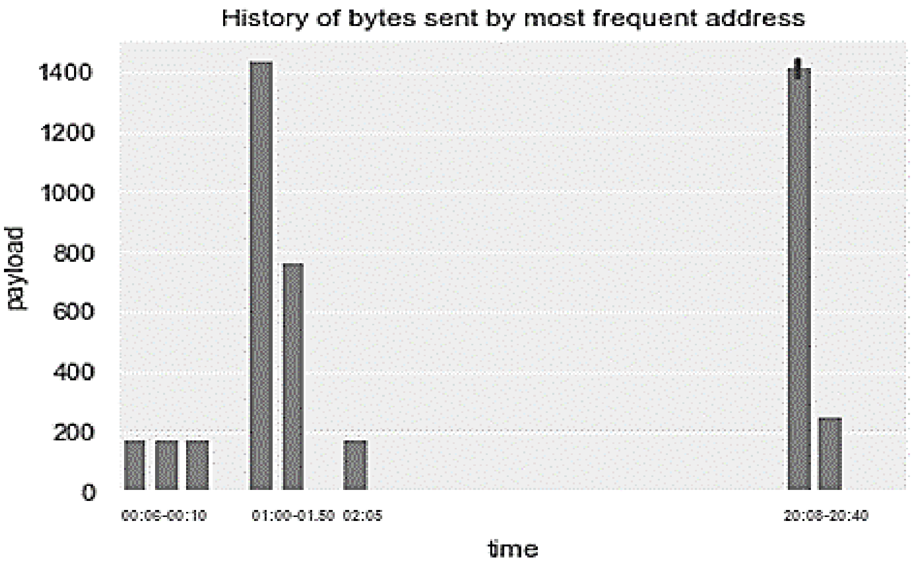


Figure 2. Size payload in data packets

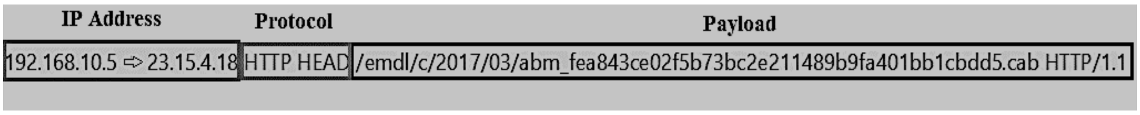
Figure 2 describes the shape of the IP payload on each communication link, which is slightly different but has a unique pattern. Three payloads are above average in size and even exceed 1500 bytes. The average payload length of a normal packet is approximately 200 bytes in size and shows no particular pattern in the payload configuration. The regularity of this iteration is analysed using the N-Gram technique with a machine learning approach. Thus, such features make it possible to find patterns and characteristics using neural networks (NNs), SVM, and KNN algorithms.

3. RESULTS AND DISCUSSION

The outcomes of building data packets using the N-Gram method are covered in this section. Normal and DDoS payloads are the two sorts of extracted payloads. First, data packets containing DDoS and regular packets from CIC-2017, MIB-2016, and H2NPayload are prepared. Next, the hex payload is extracted using online tools and the Python programming language.

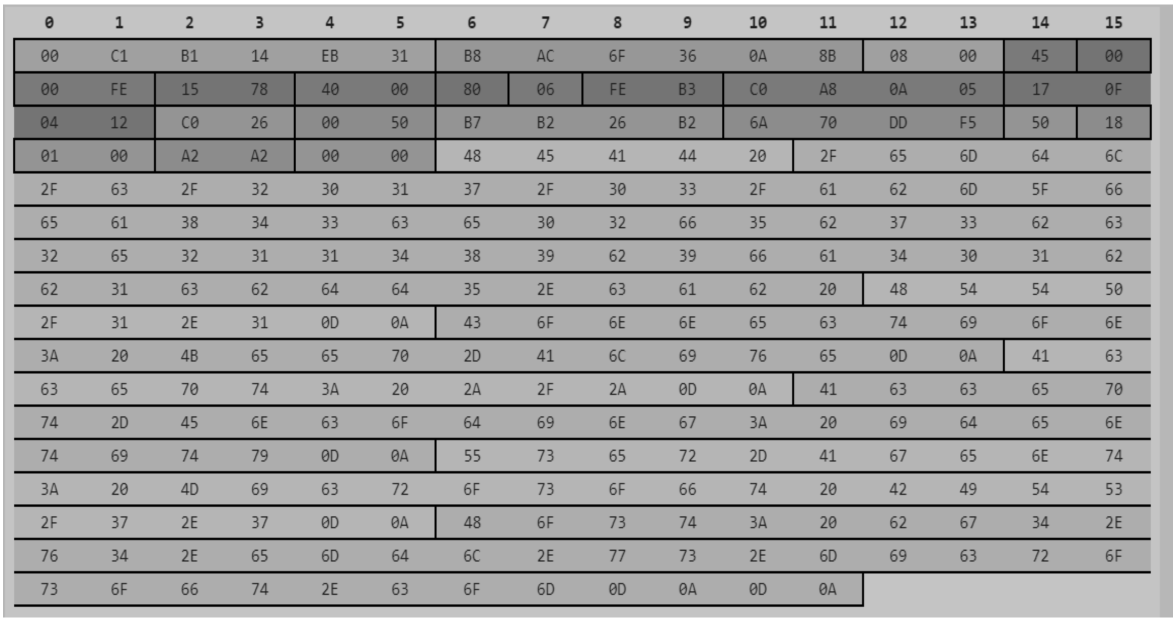
3.1. Preparation dataset result

The dataset used in this research consists of CIC-2017, CIC-2019, MIB-2016 and a new dataset called H2NPayload. Each data packet is converted from text to hexadecimal to analyse whether a packet is malicious or not. This is done because attacks are not always carried out on static features. However, attacks can be embedded into the payload, making it easier for attackers to send malicious packets to various destinations on the network, as shown in Figure 3. The outcomes of the data-gathering procedure for this investigation are shown in Figure 4. Each dataset will have every data packet thoroughly examined, with an emphasis on the payload.



IP Address	Protocol	Payload
192.168.10.5 ⇨ 23.15.4.18	HTTP HEAD	/emdl/c/2017/03/abm_fea843ce02f5b73bc2e211489b9fa401bb1cbdd5.cab HTTP/1.1

Figure 3. Payload raw



0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
00	C1	B1	14	EB	31	B8	AC	6F	36	0A	8B	08	00	45	00
00	FE	15	78	40	00	80	06	FE	B3	C0	A8	0A	05	17	0F
04	12	C0	26	00	50	B7	B2	26	B2	6A	70	DD	F5	50	18
01	00	A2	A2	00	00	48	45	41	44	20	2F	65	6D	64	6C
2F	63	2F	32	30	31	37	2F	30	33	2F	61	62	6D	5F	66
65	61	38	34	33	63	65	30	32	66	35	62	37	33	62	63
32	65	32	31	31	34	38	39	62	39	66	61	34	30	31	62
62	31	63	62	64	64	35	2E	63	61	62	20	48	54	54	50
2F	31	2E	31	0D	0A	43	6F	6E	6E	65	63	74	69	6F	6E
3A	20	4B	65	65	70	2D	41	6C	69	76	65	0D	0A	41	63
63	65	70	74	3A	20	2A	2F	2A	0D	0A	41	63	63	65	70
74	2D	45	6E	63	6F	64	69	6E	67	3A	20	69	64	65	6E
74	69	74	79	0D	0A	55	73	65	72	2D	41	67	65	6E	74
3A	20	4D	69	63	72	6F	73	6F	66	74	20	42	49	54	53
2F	37	2E	37	0D	0A	48	6F	73	74	3A	20	62	67	34	2E
76	34	2E	65	6D	64	6C	2E	77	73	2E	6D	69	63	72	6F
73	6F	66	74	2E	63	6F	6D	0D	0A	0D	0A				

Figure 4. Payload hex

3.2. Proposed N-Gram technique for DDoS attack detection

Next, employ the N-Gram approach, which is covered in the upcoming sub-chapter, to locate and rebuild the payload. When the payload from the CIC-2017 dataset data packet is extracted using the Hex Packet Decoder tool (gasmi.net), the following is what is displayed in Figure 4. The results of identifying the payload of data packets, comprising ordinary and analyzeable data packets with numerous fields, are displayed in Table 2. For every single data packet, the field descriptions shown below are applicable.

Table 2. Field packet description

Field	Hexadecimal
HTTP all	00c1b114eb31b8ac6f360a8b0800450000fe157840008006feb3c0a80a05170f0412c0260050b7b226b26a70ddf5 50180100a2a20000485414420 2f656d646c2f632f323031372f30332f61626d5f66656138343363653032663 56 2373362633265323131343839623966613430316262316362646435 2e6361622048545450 2f312e310d0a436f6 e6e656374696f6e3a204b6565702d416c6976650d0a4163636570743a202a2f2a0d0a4163636570742d456e636f 64696e673a206964656e746974790d0a557365722d4167656e743a204d6963726f736f667420424954532f372e 370d0a486f73743a206267342e76342e656d646c2e77732e6d6963726f736f66742e636f6d0d0a0d0a

Python programming was used to develop the scapy module, which separated the payload in Table 2. It is explained that the HTTP protocol's payload and data packet field are separable. Table 3 displays the outcomes of extracting and converting raw data to hexadecimal as a hex payload. Identifying the hex payload feature produces a hex payload set. The hex payload is uploaded into the pre-built tool to get the N-Gram pattern and create a new feature called CSDPayload+N-Gram, CSPayload+N-Gram, and hybrid N-Gram (CSDPayload+CSPayload N-Gram).

Table 3. Sample packet data from CIC-2017 datasets [35]

No	src	dst	sport	dport	Payload hex
0	10.1.9.1	10.51.100.44	3594	7680	b''
1	10.1.9.1	10.1.7.1	7680	57694	b''
2	10.51.100.44	10.1.9.1	7680	3594	b'000000000000'
...	...	...	...	...	...
182	23.36.33.93	192.168.10.14	80	49463	b'485454502f312e3120323030204f4b0d0a436f6e7465...
183	23.36.33.93	192.168.10.14	80	49463	b'4e616d653d224f7474617761222068696e742d6f7665...

3.3. Result N-Gram pattern formation

Determining the frequency of each payload packet string to classify data payloads into 2-Gram, 3-Gram, 4-Gram, 5-Gram, and 6-Gram categories once they have been identified and analysed for DDoS attack patterns. Once the first, second, and third datasets have all been converted, identify the payload pattern using the N-Gram approach, which ranges from 2-Gram to 6-Gram, as shown in the payload example. Table 4 shows the shift of the observed payload string and normal payload from 2-Gram to 6-Gram. String shifting aims to obtain similar and different patterns in the observed payload.

Table 4. Sliding string payload

N-Gram	Sliding string payload observed	Sliding string payload normal
2	'c1', '12', '2b', 'b0', '00', '05'…	'01', '0b', 'b1', '1c', 'c1'…
3	'c12', '12b', '2b0', 'b00'…	'00b', '0b1', 'b1b', '1c1', 'c11'…
4	'c12b', 'b005', '50a4', '4f89'…	'00b1', '0b1c', 'b1c1', '1c11'…
5	'c12b0', '0050a', 'a4f89', '96d82'…	'00b1c', '0b1c1', 'b1c11', '1c114'…
6	'c12b00', '050a4f', 'f896d', 'd8282'…	'00b1c1', '0b1c11', 'b1c114'…

3.4. Result calculation of CSD

The program will calculate the distance between the normal packet and the analysed packet using the CSD method. Calculate the pattern occurrence frequency, percentage, and CSD starting from 2-Gram, 3-Gram, 4-Gram, 5-Gram, and 6-Gram after extracting the hex payload and creating the payload string shift. The following are the processes involved in manually computing CSD using this formula:

$$D2 = \frac{(0.00332225913621262 - 0.00186915887850467)^2}{0.00332225913621262} + \frac{(0.0166112956810631 - 0.00747663551401869)^2}{0.0166112956810631}$$
$$+ \cdots \ldots + \frac{(0.0299003322259136 - 0.016822429906542)^2}{0.0299003322259136} = 0{,}327$$

Based on the following hypothesis, the analysis of the Pearson Chi-square test was performed as a threshold determination to ascertain the status of the payload observed:
$H0 : D2 \leq X2(\alpha, b-1)$
$H1 : D2 > X2(\alpha, b-1)$

$H0$ is considered a DDoS packet, but $H1$ is neither a typical payload nor a DDoS attack. The difference in the square between the two payloads is D2. The Chi-square table value, denoted as X2, has a significant value of $a = 0.05$. Its degree of freedom is $b - 1$, where b represents the number of distinct patterns found in the reference packet (normal/DDoS). Next, the value from the Chi-squared table with a value of = 0.05 and the degree of freedom $b - 1$ will be compared with the Chi-squared distance between the analysed packet and the reference packet. The Chi-squared distance computation yielded a result of 0.327. Since the value of the Chi-squared distance is less than the value of X2, the payload is a DDoS attack.

3.5. Sample feature rank generated by weight correlation

The CIC-2017 dataset comprises 78 standard features with a total of 225,745 records. This study utilises 5% of the total records for analysis. Feature selection is conducted using weight by correlation, resulting in the selection of 20 standard features with the highest correlation weight. Additionally, two new features, CSDPayload+CSPayload+N-GRAM, are added to the dataset. The CSDPayload+N-Gram feature is divided into six sub-features CSDPayload+N-Gram from 1-Gram to 6-Gram, and CSPayload+N-Gram is divided into 6 N-Gram sub-features ranging from CSPayload 1-Gram to 6 -Gram. Therefore, the total features used in this study are 32 features. Of the 32 features, the features are sorted using weight by correlation with the following calculation results:

$$\begin{aligned}
\text{Weight by correlation } (r) \quad &= \frac{S(x-\bar{x})(y-\bar{y})}{\sqrt{S(x-\bar{x})^2 S(y-\bar{y})^2}} \\
&= \frac{(92.99863561)}{\sqrt{79318.08807 * 0.14116599}} \\
&= 0.879 \\
\text{Weight by correlation (abs)} \quad &= 0.879
\end{aligned}$$

Based on the weight by correlation formula calculation results, the bwd_Packet_Lenght_Std feature has a significant relationship in determining DDoS packages with a percentage of 87.90%. The sample analysis results (Table 5) show that when selecting features in each dataset, weight by correlation is used because feature selection is calculated using polynomial data, namely a system of equations containing coefficients and variables in several terms. The weight by correlation value for each selected feature can be seen in Table 6. Table 6 shows that the Bwd_Packet_Length Std feature has a significant relationship in determining a DDoS packet or not, with a percentage of 87.90%.

Table 5. Weight by correlation value for CIC-2017 datasets

Bwd_Packet_ Length_Std (X)	Class (Y)	$x - \bar{x}$	$(x - \bar{x})^2$	$y - \bar{y}$	$(y - \bar{y})^2$	$(x - \bar{x}) * (y - \bar{y})$
0	1	-126.7956394	16,077.13	-0.1701	0.02893401	21.57
0	1	-126.7956394	16,077.13	-0.1701	0.02893401	21.57
0	1	-126.7956394	16,077.13	-0.1701	0.02893401	21.57
1130.668239	2	1003.8726	1,007,760.20	0.8299	0.68873401	833.11
635.5170373	2	508.7213979	258,797.46	0.8299	0.68873401	422.19
…	…	…	…	…	…	…
0	1	-126.7956394	16,077.13	-0.1701	0.02893401	21.57
1166.469107	2	1039.673468	1,080,920.92	0.8299	0.68873401	862.83
0	1	-126.7956394	16,077.13	-0.1701	0.02893401	21.57
0	1	-126.7956394	16,077.13	-0.1701	0.02893401	21.57
126.7956394	1.1701	0.000000009313	79318.08807	0.00000000232	0.14116599	92.99863561

Table 6. Ranking feature selection CIC-2017 dataset

No	Feature	Weight by correlation	No	Weight by correlation
1	Bwd_Packet_Length Std	0.87887	16	CS_Payload _1G
2	Bwd_Packet_Length_Max	0.81781	17	CS_Payload_6G
3	Packet_Length_Std	0.79834	18	CS_Payload _5G
4	Max_Packet_Length	0.69243	19	CS_Payload _3G
5	Bwd_Packet_Length_Mean	0.6542	20	CS_Payload _4G
6	Avg_Bwd_Segment_Size	0.6542	21	H2NPayload_2G
7	Packet_Length_Mean	0.62397	22	act_data_pkt_fwd
8	Init_Win_bytes_forward	0.45838	23	Subflow_Fwd_Packets
9	Fwd_Packet_Length_Max	0.38507	24	Subflow_Bwd_Packets
10	Avg_Fwd_Segment_Size	0.26824	25	Total_Backward_Packets
11	Down_Per_Up_Ratio	0.26173	26	Subflow_Bwd_Bytes
12	Bwd_Packet_Length_Min	0.25361	27	Total_Length_of_Bwd_Packets
13	Total_Length_of_Fwd_Packets	0.25175	28	CSD_Payload_6G
14	Min_Packet_Length	0.2153	29	CSD_Payload_5G
15	CS_Payload _2G	0.16042	30	CSD_Payload_4G

3.6. Experimentation summary

The results of experiments conducted on the four datasets in carrying out feature selection to increase the accuracy of DDoS attack detection using the hybrid N-Gram heuristic technique. Three algorithms are used to evaluate the accuracy level: SVM, KNN, and NN. The CSDPayload+N-Gram, CSPayload+N-Gram, and CSDPayload+N-Gram+CSPayload+N-Gram features are tested on the proposed model. Tables 7-9 lists the evaluation findings.

Table 7. Summary of accuracy value for four datasets using the SVM algorithm

Dataset	Features	No Features without FS/FS	Accuracy without N-Gram	N-Gram feature accuracy					
				1-G	2-G	3-G	4-G	5-G	6-G
CIC2017	CSDPayload+CSPayload+N-Gram	78/32	98.96	99.23	99.49	99.38	99.65	99.16	99.29
CIC2019	CSDPayload+CSPayload+N-Gram	77/32	97.86	99.78	99.80	99.80	99.80	99.03	99.02
MIB2016	CSDPayload+CSPayload+N-Gram	34/17	94.88	98.72	97.46	99.64	99.74	93.94	95.12
H2N-Payload	CSDPayload+CSPayload+N-Gram	6/18	58.96	98.52	98.36	98.41	99.64	97.75	98.41
Average			87.67	99.06	98.78	99.31	99.71	97.47	97.96

The 4-Gram feature is the best N-Gram size that can classify each payload. The implementation of the CSDPayload+N-Gram+CSPayload+N-Gram feature on the CIC2017, CIC2019, MIB2016 and H2NPayload datasets achieved accuracy values of 99.65%, 99.80%, 99.74% and 99.64% respectively. The average accuracy with the SVM algorithm achieved 99.71% as compared to test results without N-Gram with an 87.67% accuracy value. Thus, there was an improvement in the accuracy value of DDoS attack detection using the N-Gram technique by 12.04%. Other features also experienced a significant increase.

Table 8. Accuracy summary for four datasets using the KNN algorithm

Dataset	Features	No Features without FS/FS	Accuracy without N-Gram	N-Gram Feature Accuracy					
				1-G	2-G	3-G	4-G	5-G	6-G
CIC2017	CSDPayload+CSPayload+N-Gram	78/32	99.76	99.44	97.62	99.24	99.45	99.54	99.45
CIC2019	CSDPayload+CSPayload+N-Gram	77/32	99.57	99.70	99.70	99.70	99.71	99.70	99.70
MIB2016	CSDPayload+CSPayload+N-Gram	34/17	91.42	70.45	70.37	70.25	91.66	69.79	78.43
H2N-Payload	CSDPayload+CSPayload+N-Gram	6/18	56.24	91.97	89.00	73.15	94.06	82.91	90.69
Average			86.75	90.39	89.17	85.59	96.22	87.99	92.07

The 4-Gram feature is the best N-Gram measure that can classify each payload. Implementation of the CSDPayload+N-Gram+CSPayload+4-Gram feature on CIC2017, CIC2019, MIB2016, and H2NPayload achieved accuracy of 99.45%, 99.71%, 91.66%, 94.06%, respectively. Thus, the average accuracy with the KNN algorithm is 96.22%. Results without the N-Gram feature achieved an accuracy level of 86.75%. There was an improvement in the accuracy value of DDoS attack detection using the N-Gram technique by 9.47%.

Table 9. Accuracy summary for four datasets using the NN algorithm

Dataset	Features	No Features without FS and FS	Accuracy without N-Gram	N-Gram Feature Accuracy					
				1-G	2-G	3-G	4-G	5-G	6-G
CIC2017	CSDPayload+CSPayload+N-Gram	78/32	99.18	99.15	99.26	99.16	99.05	99.20	99.26
CIC2019	CSDPayload+CSPayload+N-Gram	77/32	99.70	99.98	99.99	99.99	99.99	99.98	99.98
MIB2016	CSDPayload+CSPayload+N-Gram	34/17	100.00	99.12	99.36	99.66	99.64	93.88	96.23
H2N-Payload	CSDPayload+CSPayload+N-Gram	6/18	57.52	98.67	99.18	99.18	99.33	98.00	96.67
Average			89.10	99.23	99.45	99.50	99.50	97.77	98.04

The 3-Gram and 4-Gram features are the best N-Gram sizes to classify each payload. The implementation of the CSDPayload+N-Gram+CSPayload+3-Gram and 4-Gram feature in the CIC2017 dataset achieved an accuracy rate of 99.05%. The CIC2019 dataset is 99.99%, the MIB2016 dataset is 99.64 %, and the H2NPayload dataset is 99.33%. Thus, the average accuracy with the NN algorithm is 99.50%. The test results without the N-Gram feature have an average accuracy rate of 89.10%. Thus, there is an increase in the accuracy of the DDoS detection value after applying the N-Gram feature by 10.40%. Table 9 also provides a detailed description of the test results before and after applying the N-Gram features. The strategies and techniques used in feature selection are also compared in this table. To highlight variations in experimental results in this study, it also analyses packet header components such as IP, TCP port, TCP flag, and payload. However, it also emphasises that a packet has several dynamic and static categories.

4. CONCLUSION

This paper suggests an N-Gram hybrid heuristic approach for DDoS attack detection. The study phase demonstrates that this technique may identify attacks by identifying the percentage of two network class circumstances (DDoS and normal) over the whole dataset. Three algorithms are used in this study: KNN, NN, and SVM. The average accuracy of the SVM algorithm for the four datasets using the CSDPayload+N-Gram feature is 99.92%, CSPayload+N-Gram is 99.72%, and the hybrid N-Gram feature is 99.71%. The KNN algorithm tested on the CSDPayload+N-Gram feature is 94.41%, CSPayload+N-Gram is 94.49%, and the on the hybrid N-Gram feature is 96.22%. While the accuracy value tested on the NN algorithm for CSDPayload+N-Gram feature is 99.88%, CSPayload+N-Gram is 99.81%, and the N-Gram hybrid feature accuracy is 99.50%.

ACKNOWLEDGEMENTS

This research was supported by Universiti Tun Hussein Onn Malaysia (UTHM) and LPPM Universitas Putera Batam.

REFERENCE

[1] S. Alzahrani and L. Hong, “Generation of DDoS Attack Dataset for Effective IDS Development and Evaluation,” *J. Inf. Secur.*, vol. 09, no. 04, pp. 225–241, 2018, doi: 10.4236/jis.2018.94016.

[2] A. Bonguet and M. Bellaiche, “A survey of Denial-of-Service and Distributed Denial of Service attacks and defences in cloud computing,” *Futur. Internet*, vol. 9, no. 3, 2017, doi: 10.3390/fi9030043.

[3] K. M. I. A. Fouda, “Payload-based signature generation for DDoS attacks.” [Online]. Available: http://essay.utwente.nl/73420/ (accesed: Aug. 23, 2017),

[4] J. Joseph and M. Dutta, “Threshold-Based Method for Detection of Distributed Denial of Service Attack in IoT,” *Int. J. Recent Technol. Eng.*, vol. 8, no. 4, pp. 3002–3007, 2019, doi: 10.35940/ijrte.d7421.118419.

[5] H. Zhao, Z. Chang, G. Bao, and X. Zeng, “Malicious Domain Names Detection Algorithm Based on N-Gram,” *J. Comput. Networks Commun.*, vol. 2019, pp. 1–9, Feb. 2019, doi: 10.1155/2019/4612474.

[6] A. Oza, K. Ross, R. M. Low, and M. Stamp, “HTTP attack detection using n-gram analysis,” *Comput. Secur.*, vol. 45, no. 2011, pp. 242–254, 2014, doi: 10.1016/j.cose.2014.06.002.

[7] K. M. Prasad, A. R. M. Reddy, and K. V. Rao, “DoS and DDoS Attacks: Defense, Detection and TracebackMechanisms -A Survey,” *Glob. J. Comput. Sci. Technol.*, vol. 14, no. 7, 2014.

[8] H. Polat, O. Polat, and A. Cetin, “Detecting DDoS attacks in software-defined networks through feature selection methods and machine learning models,” *Sustain.*, vol. 12, no. 3, 2020, doi: 10.3390/su12031035.

[9] 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.

[10] 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, doi: 10.1016/j.procs.2015.04.007.

[11] M. Revathi, V. V. Ramalingam, and B. Amutha, “A Survey of DDoS Attack Detection and Prevention Mechanism,” *J. Crit. Rev.*, vol. 7, no. 18, pp. 558–578, 2020.

[12] A. Manna and M. Alkasassbeh, “Detecting network anomalies using machine learning and SNMP-MIB dataset with IP group,” *arXiv*. 2019, doi: 10.1109/ICTCS.2019.8923043.

[13] B. B. Gupta, R. C. Joshi, and M. Misra, “ANN based scheme to predict the number of zombies in a DDoS attack,” *Int. J. Netw. Secur.*, vol. 14, no. 2, pp. 61–70, 2012.

[14] J. M. Smith and M. Schuchard, “Routing Around Congestion: Defeating DDoS Attacks and Adverse Network Conditions via Reactive BGP Routing,” in *Proceedings - IEEE Symposium on Security and Privacy*, 2018, vol. 2018-May, pp. 599–617, doi: 10.1109/SP.2018.00032.

[15] Q. Niyaz, W. Sun, and A. Y. Javaid, “A Deep Learning Based DDoS Detection System in Software-Defined Networking (SDN),” *ICST Trans. Secur. Saf.*, vol. 4, no. 12, p. 153515, 2017, doi: 10.4108/eai.28-12-2017.153515.

[16] A. Maslan, K. M. Mohammad, and S. A. Arnomo, “DDoS Detection on Network Protocol Using Cosine Similarity and N-Gram+ Method,” 2019, doi: 10.1109/SIET.2018.8693215.

[17] S. Sridharan, “Defeating N-gram Scores for HTTP Attack Detection,” *SJSU Sch. Work.*, vol. 6, pp. 1–37, 2016, doi: 10.31979/etd.japx-z6eu.

[18] M. Zekri, S. El Kafhali, N. Aboutabit, and Y. Saadi, “DDoS attack detection using machine learning techniques in cloud computing environments,” *Proc. 2017 Int. Conf. Cloud Comput. Technol. Appl. CloudTech 2017*, vol. 2018-Janua, no. February 2018, pp. 1–

7, 2018, doi: 10.1109/CloudTech.2017.8284731.

[19] M. Aldwairi, W. Mardini, and A. Alhowaide, “Anomaly payload signature generation system based on efficient tokenisation methodology,” *Int. J. Commun. Antenna Propag.*, vol. 8, no. 5, pp. 421–429, 2018, doi: 10.15866/irecap.v8i5.12794.

[20] A. Yulianto, P. Sukarno, and N. A. Suwastika, “Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset,” *J. Phys. Conf. Ser.*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012018.

[21] 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, 2021, doi: 10.1016/j.jksuci.2019.02.003.

[22] A. Shabtai, R. Moskovitch, C. Feher, S. Dolev, and Y. Elovici, “Detecting unknown malicious code by applying classification techniques on OpCode patterns,” *Secur. Inform.*, vol. 1, no. 1, p. 1, 2012, doi: 10.1186/2190-8532-1-1.

[23] N. Bindra and M. Sood, “Evaluating the impact of feature selection methods on the performance of the machine learning models in detecting DDoS attacks,” *Rom. J. Inf. Sci. Technol.*, vol. 23, no. 3, pp. 250–261, 2020.

[24] D. Stiawan, S. M. Daely, A. Heryanto, N. Afifah, M. Y. Idris, and R. Budiarto, “Ransomware detection based on opcode behaviour using k-nearest neighbours algorithm,” *Inf. Technol. Control*, vol. 50, no. 3, pp. 495–506, 2021, doi: 10.5755/j01.itc.50.3.25816.

[25] K. Swapna, M. C. B. Prasad, “Semi-Supervised Machine Learning For Ddos Attack Classification Using Clustering Based,” *Journal of Engineering Sciences*, vol. 12, no. 12, pp. 472–478, 2021.

[26] L. Csikar, “Decision making in the sciences : understanding heuristic use by students in problem solving,” Ph.D. dissertation, Instruc. Des. Technol., *Old Dominion University*, 2018.

[27] S. Khunkitti, A. Siritaratiwat, and S. Premrudeepreechacharn, “Multi-objective optimal power flow problems based on slime mould algorithm,” *Sustain.*, vol. 13, no. 13, 2021, doi: 10.3390/su13137448.

[28] S. Khunkitti, A. Siritaratiwat, and S. Premrudeepreechacharn, “A Many-Objective Marine Predators Algorithm for Solving Many-Objective Optimal Power Flow Problem,” *Appl. Sci.*, vol. 12, no. 22, 2022, doi: 10.3390/app122211829.

[29] Canadian Institute for Cybersecurity, “Dataset CIC-2017,” 2017. [Online]. Available: https://www.unb.ca/cic/datasets/ids-2017.html. (accessed Jun. 17, 2018).

[30] C. Ma, X. Du, and L. Cao, “Analysis of multi-Types of flow features based on hybrid neural network for improving network anomaly detection,” *IEEE Access*, vol. 7, pp. 148363–148380, 2019, doi: 10.1109/ACCESS.2019.2946708.

[31] 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, doi: 10.14569/IJACSA.2016.070159.

[32] Z. Chiba, N. Abghour, K. Moussaid, A. El, and M. Rida, “Intelligent and Improved Self-Adaptive Anomaly based Intrusion Detection System for Networks,” *Int. J. Commun. Net. Inf. Secur.*, vol. 11, no. 2, pp. 312–330, 2019.

[33] K. Kumari and M. Mrunalini, “Detecting Denial of Service attacks using machine learning algorithms,” *J. Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00616-0.

[34] S. Brindha, M. P. Abirami, V. Arjun, B. Logesh, and M. S. P, “Heuristic Approach to Intrusion Detection System,” *Int. Res. J. Eng. Technol.* vol. 7, no. 3, pp. 377–379, 2020.

[35] A. Maslan, K. M. Mohamad, and C. F. M. Foozy, “Enhancement detection distributed denial of service attacks using hybrid n-gram techniques,” *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. 20, no. 1, pp. 61–69, 2022, doi: 10.12928/TELKOMNIKA.v20i1.18103.

BIOGRAPHIES OF AUTHORS

Andi Maslan received a degree in Informatics Engineering at the Budi Utomo Institute of Technology Jakarta (2004) and a Master’s degree in computer science (Information Systems) at the STMIK Putera Batam (2011). Currently, he is finishing his doctoral education at Tun Onn Hussein University Malaysia. He is a lecturer at the University of Putera Batam and has a functional position as an assistant professor. His current research interests include networking, network security, and artificial intelligence. He can be contacted at email: Lanmasco@gmail.com.

Abdul Hamid received a Ph.D. in Engineering Technology from the Universiti Tun Hussein Onn Malaysia (UTHM) in 2019. He is currently a senior lecturer with the Department of Technology Studies, UTHM Johor Malaysia. He has published 40 academic papers as a first author or a co-author in conference proceedings and international journals. His research interests in applied sciences, engineering and technologies include smart manufacturing, transportation and society, mechanical engineering, informatics visualisation, and geoinformatics. He can be contacted at email: abdulhamid@uthm.edu.my.

Dedy Fitriawan received a Master of Science in Geography Sciences from Universitas Indonesia (UI) in 2014. He is currently a junior lecturer at the Department of Remote Sensing and Geographic Information Systems (RSGIS), School of Vocational Universitas Negeri Padang (UNP) Padang, Indonesia. He published 10 scientific papers as first author/co-author in conference proceedings and national/international journals. His research interests include applied geosciences, remote sensing/photogrammetry and satellite/aerial image processing, applied computational geosciences with AI, and geoinformatics. He can be contacted at email: dedyfitriawan@unp.ac.id

Anggia Dasa Putri is a lecturer in the Informatics Engineering Study Program at Putera University Batam. She graduated from Universitas Putra Indonesia "YPTK" Padang, earning a Bachelor's degree in 2012 and a Master's in Computer Science (Information Systems) in 2014. Currently, she is pursuing her Doctoral education at Universitas Putra Indonesia "YPTK" Padang. She holds a position as a lecturer at Universitas Putera Batam and is actively engaged in research, with interests focusing on artificial intelligence and big data. She can be contacted at email: anggiaputri4@gmail.com.

Tukino is a lecturer in the Information Systems Study Program at Putera University Batam. He graduated from STMIK Putera Batam, earning a Bachelor's degree in 2010 and a Master's in Computer Science (Information Systems) in 2012. Currently, he is pursuing his Doctoral education at Universitas Putra Indonesia "YPTK" Padang. He holds a position as a lecturer at Universitas Putera Batam and is actively engaged in research, with interests focusing on artificial intelligence and big data. He can be contacted at email: tukino@puterabatam.ac.id.