

Hybrid N-gram-based framework for payload distributed denial of service detection and classification

Andi Maslan¹, Cik Feresa Mohd Foozy², Kamaruddin Malik Mohamad², Abdul Hamid³,
Dedy Fitriawan⁴, Joni Hasugian⁵

¹Department of Informatic Engineering, Faculty of Engineering and Computer Science, Universitas Putera Batam, Batam, Indonesia

²Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

³Faculty of Technical and Vocational Education, Universiti Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

⁴Department of Remote Sensing and Geographic Information System, Vocational School, Universitas Negeri Padang, Padang, Indonesia

⁵PT. Angkasa Pura Indonesia, Jakarta, Indonesia

Article Info

Article history:

Received Jun 26, 2024

Revised Sep 29, 2025

Accepted Oct 16, 2025

Keywords:

Chi square

Cosine similarity

DDoS

Network

Payload

ABSTRACT

There are three main approaches to distributed denial of service (DDoS) detection: anomaly-based, pattern-based, and heuristic-based. The heuristic-based approach combines the strengths of both anomaly and pattern detection. However, existing DDoS detection systems still struggle with hypertext transfer protocol (HTTP) payload-level analysis due to high false positive rates and limited dataset granularity. To overcome these limitations, this study proposes a novel heuristic method based on a hybrid N-gram model that integrates two key components: chi-square distance (CSD)Payload+N-gram and cosine similarity (CS)Payload+N-gram. The CSDPayload measures the difference between a given payload and normal traffic using the CSD, while CSPayload evaluates their similarity using CS. These metrics form a comprehensive feature set evaluated on three benchmark datasets: CIC2019, MIB2016, and H2N-Payload. The methodology involves extracting HTTP traffic, converting it into hexadecimal payloads, and applying N-gram analysis (1- to 6-Gram). Frequency distributions are used to calculate CSD, CS, and Pearson's chi-square test for payload classification. Feature selection based on weight correlation refines the input for machine learning classifiers support vector machine (SVM), k-nearest neighbors (KNN), and neural network (NN). Experimental results indicate high accuracy, particularly for the 4-Gram model: NN achieves 99.65%, KNN 95.14%, and SVM 99.73%.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Andi Maslan

Department of Informatic Engineering, Faculty of Engineering and Computer Science

Universitas Putera Batam

Tembesi, Kepulauan Riau, Batam, Indonesia

Email: Lanmasco@gmail.com

1. INTRODUCTION

Distributed denial of service (DDoS) attacks has persisted for decades. In the past, such attacks often originated from a limited number of sources, which could be effectively mitigated using specific defense mechanisms—typically by blocking or denying access from those identified sources, especially when enhanced traceability was available. However, with the exponential growth of the internet, modern systems have become increasingly vulnerable. The sheer volume of simultaneous data access requests directed at servers is now difficult to manage, creating opportunities for attackers to overwhelm or bypass server defenses, whether deliberately or inadvertently.

Three methods are often employed in detecting DDoS attacks: DDoS detection approaches are generally categorized into three types: pattern-based, anomaly-based, and heuristic-based methods. Each of these techniques has its own strengths and limitations, meaning that no single method offers a universally optimal or definitive solution. Specifically, pattern-based detection works by comparing sequences of data packets traversing a network against a predefined set of rules or known malicious payload patterns. This method is quite powerful for previously recognized attacks, which are still often used today [1]. The disadvantage of this pattern-based method is that if the incoming attack has never existed or the list of rules used has not changed for too long, it cannot detect the latest attacks. Then, the detection accuracy level is still low, namely 95%, which can still be improved [2].

Then, Aldwairi *et al.* [3] proposed a DDoS attack detection method based on payload similarity, employing a similarity-based classification approach. In a related study, Masud *et al.* [4] introduced a feature selection technique using information gain to enhance detection accuracy, combined with a random forest classifier that achieved a detection accuracy of 99.06% and a low false alarm rate of 0.094. Meanwhile, Zahid and Bharati [5] presented a hybrid approach to process streaming network packets and classify DDoS attacks, reporting hybrid deep learning model (CNN-BiLSTM) an accuracy of 99.9%. Furthermore, Masud *et al.* [4] highlighted that intrusion detection systems (IDS) are effective for attack detection, as they can identify suspicious behavior, anomalous traffic patterns, and even previously unknown attack types—particularly those exploiting synchronize (SYN) packets, which allow the system to detect discrepancies between data packets transmitted over the network.

Additionally, Bindra and Sood [6] investigated the influence of feature selection on the performance of machine learning models in DDoS detection. It concluded that signature-based defense mechanisms are inadequate against evolving threats like DDoS attacks. As emphasized in [7], the core objective of developing a machine learning classifier is to detect DDoS attacks both efficiently and effectively. However, model performance heavily depends on the selection of relevant features from network traffic. The research evaluated 84 standard features using several machine learning classifiers—including support vector machine (SVM), Gaussian naive Bayes (GNB), k-nearest neighbors (KNN), and random forest. Among these, KNN achieved 94% accuracy with 15-fold cross-validation, while random forest yielded the highest performance at 96% accuracy.

At the moment, DDoS is a kind of cyberattack that can target any website, including those operated by businesses, schools, individuals, and online retailers. The attacks also keep changing in tandem with technological advancements. Layers two through seven are the attack target since this is where the server loads the webpage and responds to hypertext transfer protocol (HTTP) requests. Because it mimics real online traffic, this type of attack is often hard to recognize and counter.

Kim *et al.* [8] have expressed ongoing concerns regarding traffic analysis methods that rely solely on statistical metrics—such as packet count, size, and transmission duration. Traditionally, DDoS detection involves aggregating individual packets into network flows based on the five-tuple: source IP, source port, destination IP, destination port, and transport-layer protocol. However, HTTP-based DDoS attacks have received comparatively less attention because their detection often requires inspecting payload content, which is only accessible after flow completion. This introduces additional computational overhead when extracting statistical features from flows.

While existing methods can identify both bandwidth- and resource-depletion DDoS attacks, most focus primarily on bandwidth-related indicators—such as the volume and size of incoming/outgoing packets—leading to high false positive rates. To address these limitations, [6] proposed approaches leveraging statistical analysis of datasets from the management information base (MIB) and the Canadian Institute for Cybersecurity (CIC). More recently, other studies [9]–[11] have employed machine learning techniques to detect network intrusions and anomalies. In particular, Wang *et al.* [12] conducted a comprehensive review of anomaly detection methods using the MIB2016 and CIC2017 datasets, while Manna and Alkasassbeh [13] introduced a new dataset incorporating modern attack vectors not previously covered in the literature. Their methodology utilizes 91 MIB-derived traffic features grouped into five protocol categories: internet protocol (IP), internet control message protocol (ICMP), transmission control protocol (TCP), user datagram protocol (UDP), and simple network management protocol (SNMP); collected periodically from both attack sources and target systems. The experimental setup included three controlled DDoS attack types: Ping Flood, Targa3, and UDP Flood [9].

Despite these advances, pattern-based detection in IDS faces two key challenges. First, DDoS attacks are relatively easy to launch and difficult to trace due to inherent limitations in the TCP/IP protocol suite, which attackers exploit to obscure victim identification [14]. Moreover, modern DDoS tactics—such as SYN-Flood attacks—further complicate detection. A single SYN packet is typically indistinguishable from legitimate traffic, making it hard for IDS to flag such activity as anomalous. Consequently, SYN-Flood attacks often evade early warning systems. Second, signature-based IDS frequently generate false positives

when normal network behavior is misclassified as malicious [2]. Given these challenges, timely detection and rapid deployment of mitigation strategies are critical to preserving network availability and functionality during DDoS incidents.

Then, Swapna and Prasad [15] proposed a study using the N-gram method to examine the network traffic flow header. Selection of the best features using the chi-square test aims to know the significant relationship between the two variables being compared. At the same time, determining the order of features using an algorithm based on the order of N-gram to get meaningful features from the semantics of traffic flow. In addition to detecting malware, this method can also detect DDoS attacks that focus on the HTTP protocol, whether web-based or attacks on mobile networks. The results demonstrate the solution's efficiency, and a trained model can identify malicious attacks with multiple false warnings. The detection accuracy rate is 99.15%, but the false positive is 0.45%. It can detect 54.81% of malicious applications when used in a real environment, which is better than other popular anti-virus scanners.

From the background of the problem and the motivation of the research, N-gram-based payload-level detection is an approach that utilizes the analysis of the payload content of network packets to detect DDoS attack patterns, especially at the application layer (layer 7). This technique performs feature extraction based on a sequence of characters or bytes (N-gram) in a payload. Which is then used to distinguish between normal traffic and offensive traffic.

Furthermore, N-gram is heuristic-based, where the process of tokenizing the payload into N-gram is combined with rules or patterns designed based on domain knowledge and malicious traffic characteristics. Heuristics are used to filter and emphasize specific N-grams that have high relevance to attack behavior, so that they can improve detection efficiency and accuracy without relying entirely on statistical learning methods or complex classification models. While the third stage, heuristic and hybrid techniques, is an approach that combines rule-based methods with machine learning techniques to increase detection effectiveness. Heuristic techniques in this context refer to the use of domain knowledge and known attack behavior patterns to form initial detection rules, such as recognizing abnormal frequencies of a particular N-gram or payload structures that deviate from normal traffic.

Meanwhile, the hybrid approach integrates heuristics with learning algorithms, both sequentially (heuristic as a pre-processing stage before classification) and parallel (the results of heuristics and statistical models are combined for decision-making). This combination aims to overcome the limitations of each method, where heuristics excel in quick detection and explicit knowledge-based, while machine learning has the advantage of capturing complex patterns and generalizing to new attack variations. By utilizing heuristic and hybrid techniques, the detection system is able to carry out early identification of suspicious traffic, while increasing accuracy rates and lowering false positive rates. This strategy is particularly relevant in dynamic DDoS attack scenarios, where attack patterns can be fickle and difficult to detect with a single approach.

2. METHOD

This study employs a heuristic-based N-gram technique for DDoS attack detection. The research begins with the collection of three datasets: CIC2019 [16], [17], MIB2016 [18], and a newly generated dataset derived from a simulated DDoS attack using a custom-built tool named Hammer Master, implemented in a programming language (referred to as H2N-Payload). The composition of these datasets is summarized in Table 1.

Table 1. Dataset details

No	Dataset	Total samples	Payload size	DDoS	Normal
1	CIC-2019	10,000	291 bytes	8,316	1,684
2	MIB-2016	4,998	196 bytes	3,105	6,895
3	H2N-Payload	1,954	89 bytes	1,094	860

After collecting data, proceed to the second stage by proposing a construction model for DDoS detection by identifying the payload using the online application hpd.gasmi.net and the scapy module run through Jupyter Notebook. Raw data is uploaded, then the general fields are separated into packet data. The fields include Ethernet, IPv4, TCP, and HTTP0.

At this stage, payload identification focuses on the HTTP protocol [19]. After separating the fields, the data packet payloads are collected. An analysis is carried out to the next stage, such as forming a pattern by taking a hexadecimal string from the normal payload, then taking the string pattern from the observed payload and storing the patterns that appear, calculating the frequency and total frequency of each. Each pattern starts

from 1- to 6-Gram, calculates chi-square distance (CSD) between packets, and performs normal packet classification or DDoS attacks using machine learning. The determination of whether the package is dangerous or not is determined based on the Pearson chi-square test analysis results, according to the hypothesis formed [20]. After this stage is done, all payloads analyzed will be labeled normal or DDoS classes.

Then in the third stage, implementing hybrid N-gram [21]. Heuristic techniques by selecting features on the dataset formed from 1-Gram to 6-Gram, feature selection using weight correlation. After selecting the features, the payload classification is done using the SVM, KNN, and NN algorithms. In the final stage, the three datasets enhanced with the newly extracted features are evaluated using standard performance metrics, including accuracy, precision, recall, F-measure, and receiver operating characteristic-area under the curve (ROC-AUC), to assess their effectiveness in detecting DDoS attacks. A comparative analysis of the performance across different machines learning algorithms is then conducted. The overall research methodology [22] is illustrated in Figure 1.

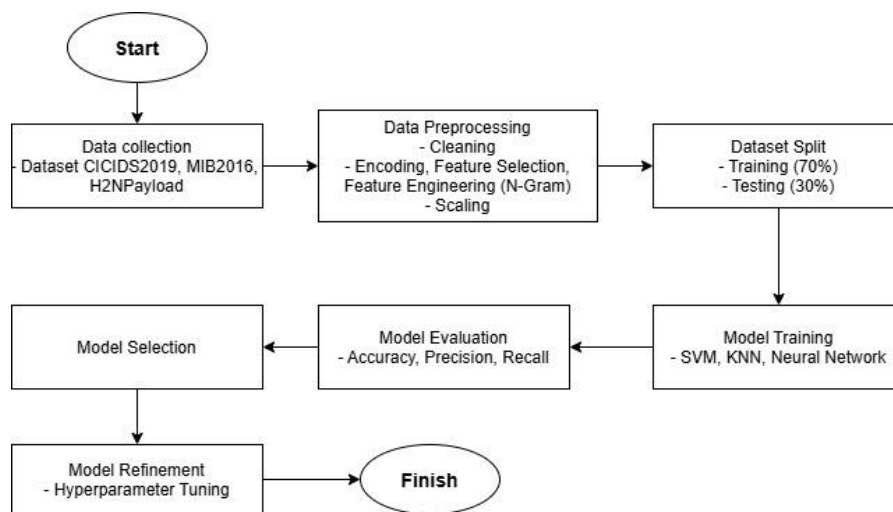


Figure 1. AI-based design and experimental methods applied

Figure 1 outlines the proposed DDoS attack detection approach, which is based on a heuristic framework. This framework is generally divided into two complementary components: pattern-based and anomaly-based detection—either or both of which may be employed depending on the context. The heuristic-based methodology encompasses four main categories of DDoS detection techniques: knowledge-based, statistical-based, soft computing-based, and machine learning-based. Each category employs distinct algorithms tailored to its underlying principles. When a knowledge-based approach is adopted, the focus shifts to analyzing packet structures, particularly headers and payloads. In contrast, a statistical-based approach leverages models such as CSD, correlation analysis, analysis of variance (ANOVA), and both parametric and non-parametric statistical tests to identify deviations from normal traffic behavior.

The soft computing-based category integrates highly efficient algorithms and advanced computational techniques, including fuzzy logic, artificial neural networks, and probabilistic reasoning to handle the uncertainty and complexity inherent in network traffic [23]. Similarly, machine learning serves as an intelligent system capable of improving its performance over time through experience, adapting to new attack patterns based on feedback from prior tasks and evaluation metrics [24]. Notably, heuristic methods in this study perform deep inspection of the HTTP protocol, specifically examining packet contents associated with common request methods such as POST, GET, and other protocol-specific commands [25]. The payload is first extracted and converted into hexadecimal format to enable N-gram analysis. Two key similarity measures are then applied: CSD and cosine similarity (CS). CSD quantifies the divergence between an observed payload and a baseline (normal) payload, while CS measures their degree of similarity where a CS value closer to 1 indicates higher resemblance to normal traffic.

These computations yield two novel hybrid features: CSDPayload+N-gram and CSPayload+N-gram. Each feature is assigned a numerical value and a decision threshold, which together determine whether a given packet should be classified as malicious. The pseudocode or algorithmic formulation used to derive this hybrid N-gram feature (combining CSDPayload and CSPayload) is presented in Pseudocode 1.

Pseudocode 1. Process of creating the N-gram patterns

Aim: Making N-gram Patterns

Input: Value of n and N-gram

Output: pattern as N-gram that appears

```

def StringBreaker(string,divider):
    i=0
    result={}
    mylist=[]
    str_len=len(string)
    while(i<str_len):
        newstr=string[i:i+divider]
        new_str=newstr.replace(' ','space') # replace ' '
        mylist.append(new_str)
        i+=1
    result['original_string']=string
    result['char_separation']=mylist

    dict={}
    for n in mylist:
        keys=dict.keys()
        if n in keys: #
            dict[n] +=1
        else:
            dict[n]=1
    result['char_grouping']=dict
    return result

```

Pseudocode describes the process of creating the N-gram patterns that appear in the packet as follows:

- i) Create a function to split the string.
- ii) Declare the result, which is a value return object.
- iii) Declare array my list to hold string values split per divider.
- iv) Calculating string length
- v) If less than string length, then push data to array mylist.
- vi) In each character, take and read the character along the divider.
- vii) Use the word space to make it easier.
- viii) Add the word space to mylist.
- ix) Declare a variable dict to store calculation results.
- x) Specify alias variable for each value in mylist.
- xi) If the variable has been read in the previous value, then set qty+1 for the variable

To implement this algorithm, the program modules used in python programming are Jupyter Notebook and scikit-learn.

3. RESULTS AND DISCUSSION

This section presents the results of data packet construction using the N-gram method. The extracted payloads are categorized into two types: DDoS payloads and normal (benign) payloads. Initially, network traffic data containing both DDoS attack packets [26] and regular traffic packets are collected from three sources: the CIC2019, MIB2016, and H2N-Payload datasets. Subsequently, the payload portions of these packets are extracted in hexadecimal format using a combination of online tools and custom scripts developed in Python.

3.1. Preparation dataset result

The identified payloads [26] were extracted from unprocessed data for further examination. Before being converted into hexadecimal form, the raw data was identified as a packet capture (PCAP) file [27] from the CIC-2019 dataset, then extracted using the Scapy Python module as shown in Figure 2. The first step in analyzing the data payload is to identify raw data in the packet. It can be seen that the blue one is a feature of the data packet, while the red one is the protocol type and also describes the open systems interconnection (OSI) layer. Then, the conversion process from text to hexadecimal is carried out as shown in Figure 3.

3.2. Proposed N-gram technique for DDoS attacks detection

Step 2 employs the N-gram approach detailed in the following sub-chapter to locate and reconstruct the payload. When the payload from a data packet in the CIC-2019 dataset is extracted using the Hex Packet Decoder tool (gasmi.net), the resulting output is displayed in Figure 4. Figure 4 shows the outcome of

[illegible]

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

192.168.50.6 → 23.63.78.40 HTTP GET /success.txt HTTP/1.1															
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
70	F3	5A	42	73	E8	B8	CA	3A	7B	97	D6	08	00	45	00
01	51	74	D1	40	00	80	06	2C	C0	C0	A8	32	06	17	3F
4E	28	D6	1F	00	50	25	99	EF	0F	23	C5	E9	6E	50	18
01	00	7B	81	00	00	47	45	54	20	2F	73	75	63	63	65
73	73	2E	74	78	74	20	48	54	54	50	2F	31	2E	31	0D
0A	48	6F	73	74	3A	20	64	65	74	65	63	74	70	6F	72
74	61	6C	2E	66	69	72	65	66	6F	78	2E	63	6F	6D	0D
0A	55	73	65	72	2D	41	67	65	6E	74	3A	20	4D	6F	7A
69	6C	6C	61	2F	35	2E	30	20	28	57	69	6E	64	6F	77
73	20	4E	54	20	36	2E	33	3B	20	57	69	6E	36	34	3B
20	78	36	34	3B	20	72	76	3A	36	33	2E	30	29	20	47
65	63	6B	6F	2F	32	30	31	30	30	31	30	31	20	46	69
72	65	66	6F	78	2F	36	33	2E	30	0D	0A	41	63	63	65
70	74	3A	20	2A	2F	2A	0D	0A	41	63	63	65	70	74	2D
4C	61	6E	67	75	61	67	65	3A	20	65	6E	2D	55	53	2C
65	6E	3B	71	3D	30	2E	35	0D	0A	41	63	63	65	70	74
2D	45	6E	63	6F	64	69	6E	67	3A	20	67	7A	69	70	2C
20	64	65	66	6C	61	74	65	0D	0A	43	61	63	68	65	2D

Payload separation, as shown in Table 2, was performed using the Scapy module implemented in Python. This process demonstrates that the data packet fields, and the HTTP protocol payload can be effectively isolated. Table 3 presents the results of extracting raw network data and converting it into hexadecimal format, yielding the hex payload used for subsequent analysis.

Table 2. Field packet description

Field	Hexadecimal
Ethernet	00c1b114eb31b8ac6f360a8b0800
IPv4	4500016c352a40008006e80dc0a80a0f0d6b0432
TCP	c12b0050a4f896d828237ace5018010249620000
HTTP	00c1b114eb31b8ac6f360a8b0800450000fe157840008006feb3c0a80a05170f0412c0260050b7b226b26a70ddf550180100a2a2000048454144202f656d646c2f632f323031372f30332f61626d5f666561383433636530326635623733626332653231313438396239666134303162623163626464352e63616220485454502f312e310d0a436f6e6e656374696f6e3a204b6565702d416c6976650d0a4163636570743a202a2f2a0d0a4163636570742d456e636f64696e673a206964656e746974790d0a557365722d4167656e743a204d6963726f736f667420424954532f372e370d0a486f73743a206267342e76342e656d646c2e77732e6d6963726f736f66742e636f6d0d0a0d0a

Table 3. Sample packet data from CIC-2019 datasets

No	src	dst	sport	dport	Payload hex
1	11.51.100.45	10.1.9.1	7680	3594	b'000000000000'
2	23.36.33.93	192.168.10.14	80	49463	b'485454502f312e3120323030204f4b0d0a436f6e7465...'
3	23.36.33.93	192.168.10.14	80	49463	b'4e616d653d224f7474617761222068696e742d6f7665...'
184	192.168.10.5	23.15.4.18	49190	80	b'0a486f73743a206267342e76342e656d646c2e77732e6d6963726f736f66742e636f6d0d0a0d0a...'
185	23.194.182.63	192.168.10.14	80	49462	b'486f73743a2061752e646f776e6c6f61642e77696e646f7773750646174652e636f6d0d0a0d0a...'

3.3. Result N-gram pattern formation

After identifying and analyzing payload strings for DDoS attack patterns, the frequency of each N-gram sequence is calculated to categorize the payloads into 2-, 3-, 4-, 5-, and 6-Gram groups. Once all three datasets have been processed and converted, the N-gram approach spanning from 2- to 6-Gram is applied to identify recurring payload patterns. An illustrative example of this payload analysis is provided in Table 4. Based on the N-gram pattern formation model in Figure 4, Table 4 shows the shift of observed charge and normal charge from 2- to 6-Gram. Examples of observed charge and normal charge for 2-, 3-, 4-, 5- and 6-Ngram are explained in Table 4 which shows the formation of N-gram pattern.

Table 4. Sliding string payload

N-gram	Sliding string payload observed	Sliding string payload normal
2	'00', '0c', 'c1', '1b', 'b1', '11'...	'00', '0c', 'c1', '1b', 'b1'...
3	'00c', '0c1', 'c1b', '1b1'...	'00c', '0c1', 'c1b', '1b1', 'b11'...
4	'00c1', '0c1b', 'c1b1', '1b11'...	'00c1', '0c1b', 'c1b1', '1b11'...
5	'00c1b', '0c1b1', 'c1b11', '1b114'...	'00c1b', '0c1b1', 'c1b11', '1b114'...
6	'00c1b1', '0c1b11', 'c1b114', '1b114e'...	'00c1b1', '0c1b11', 'c1b114'...

3.4. Result calculation of chi-square distance

This technique quantifies the divergence between normal (benign) packets and packets analyzed using the CSD method. After extracting the hexadecimal payload and generating a shifted payload sequence, the software calculates the frequency, relative percentage, and CSD for each N-gram pattern, specifically for 2- to 6-Gram. Manual CSD calculation using this algorithm is performed through (1).

$$D2 = \frac{(0.00332225913621262 - 0.00186915887850467)^2}{0.00332225913621262} + \frac{(0.0166112956810631 - 0.00747663551401869)^2}{0.0166112956810631} + \dots + \frac{(0.0299003322259136 - 0.016822429906542)^2}{0.0299003322259136} = 0.327 \quad (1)$$

The Pearson chi-square test was applied to determine an appropriate threshold for classifying the observed payload, based on the hypotheses as in (2) and (3).

$$\text{Null hypothesis } (H_0): D2 \leq \chi^2(\alpha, b - 1) \quad (2)$$

$$\text{Alternative hypothesis } (H_1): D2 > \chi^2(\alpha, b - 1) \quad (3)$$

Here, D_2 represents the CSD between the analyzed payload and a reference payload (either normal or DDoS), b denotes the number of distinct N-gram patterns in the reference payload, and the degrees of freedom are $b-1$. The significance level is set at $\alpha=0.05$.

In this context, H_0 indicates that the payload is consistent with a DDoS attack (i.e., it does not significantly differ from the DDoS reference), whereas H_1 suggests the payload is neither typical benign traffic nor a known DDoS pattern (i.e., it exhibits statistically significant deviation). The analysis compared the computed CSD ($D_2=0.327$) against the critical value from the chi-square distribution table: $\chi^2_{(0.05,146)}=176.293$. Since $0.327 < 176.293$, the null hypothesis (H_0) is not rejected, leading to the conclusion that the payload is classified as a DDoS attack.

The results of the CSD calculation for 2-grams are shown in Table 5, which indicates that there is a variation in the patterns observed within the packets. Therefore, the patterns found in the observed payload should be used to calculate the CSD value. Consequently, the frequency of a pattern in the analyzed packet is considered zero if it appears in the observed payload but not in the studied packet.

Table 5. 2-Gram payload pattern

No	Observed payload	F	Normal payload	F
1	49	1	71	1
2	0c	1	7d	1
3	c1	1	c1	1
4	40	1	da	1
5	b1	1	b1	1
6	11	1	11	1
7	3c	1	14	1
13	8a	1	8a	1
14	ac	1	c7	1
15	86	1	49	1
16	4d	1	34	1
17	a5	1	f1	1
18	0a	1	1d	1
20	90	1	18	1
21	51	1	53	1
22	8b	1	35	1
23	b0	1	09	1
147	63	16		0

3.5. Experimentation summary

The experiments were conducted on four datasets to evaluate the effectiveness of feature selection in improving the accuracy of DDoS attack detection using the proposed hybrid N-gram heuristic technique. Three machine learning algorithms SVM, KNN, and NN sets: CSDPayload+N-gram, CSPayload+N-gram, and the combined hybrid N-gram (CSDPayload+CSPayload+N-gram). The detailed evaluation results are summarized in Tables 6 through 8.

The 4-Gram configuration emerged as the optimal N-gram size for payload classification. When the hybrid feature CSDPayload+N-gram+CSPayload+N-gram was applied to the CIC-2019, MIB-2016, and H2N-Payload datasets, it achieved detection accuracies of 99.80, 99.74, and 99.64%, respectively. Using the SVM algorithm, the average accuracy across the three datasets reached 99.73%, a substantial improvement over the baseline model without N-gram features, which yielded only 83.90% accuracy. This represents an absolute accuracy gain of 15.83% points (not 12.73%, as $99.73\%-83.90\%=15.83\%$), demonstrating the significant enhancement in DDoS detection performance enabled by the proposed N-gram technique. Other feature variants also showed notable improvements in classification accuracy.

The 4-Gram configuration proved to be the most effective N-gram size for payload classification. When the hybrid feature CSDPayload+N-gram+CSPayload+4-Gram was applied to the CIC-2019, MIB-2016, and H2N-Payload datasets using the KNN algorithm, it achieved classification accuracies of 99.71, 91.66, and 94.06%, respectively. This yields an average accuracy of 95.14%. In comparison, the same model without the N-gram feature achieved only 82.41% accuracy. The incorporation of the N-gram technique thus improved detection accuracy by 12.73% points, highlighting its effectiveness in enhancing DDoS attack detection performance.

The 3- and 4-Gram configurations emerged as the most effective N-gram sizes for payload classification. When the hybrid features CSDPayload+N-gram+CSPayload+3-Gram and CSDPayload+N-gram+CSPayload+4-Gram were applied using a NN classifier, they achieved high detection accuracies across the three datasets: 99.99% on CIC-2019, 99.64% on MIB-2016, and 99.33% on H2N-Payload. This results in

an average accuracy of 99.65%. In contrast, the same model without the N-gram features achieved an average accuracy of only 85.74%. The integration of the N-gram technique therefore improved DDoS detection accuracy by 13.91% points ($99.65\% - 85.74\% = 13.91\%$), underscoring its significant contribution to detection performance.

Table 6. Accuracy detail for four datasets using the SVM algorithm

Dataset	Features	Accuracy without	N-gram feature accuracy					
		N-gram	1-G	2-G	3-G	4-G	5-G	6-G
CIC2019	CSDPayload+CSPayload+N-gram	97.86	99.78	99.80	99.80	99.80	99.03	99.02
MIB2016	CSDPayload+CSPayload+N-gram	94.88	98.72	97.46	99.64	99.74	93.94	95.12
H2N-Payload	CSDPayload+CSPayload+N-gram	58.96	98.52	98.36	98.41	99.64	97.75	98.41
Average		83.90	99.01	98.54	99.28	99.73	96.91	83.90

Table 7. Accuracy detail for four datasets using the KNN algorithm

Dataset	Features	Accuracy without	N-gram feature accuracy					
		N-gram	1-G	2-G	3-G	4-G	5-G	6-G
CIC2019	CSDPayload+CSPayload+N-gram	99.57	99.70	99.70	99.70	99.71	99.70	99.70
MIB2016	CSDPayload+CSPayload+N-gram	91.42	70.45	70.37	70.25	91.66	69.79	78.43
H2N-Payload	CSDPayload+CSPayload+N-gram	56.24	91.97	89.00	73.15	94.06	82.91	90.69
Average		82.41	87.37	86.36	81.03	95.14	84.13	89.61

Table 8. Accuracy detail for four datasets using the neural network algorithm

Dataset	Features	Accuracy without	N-gram feature accuracy					
		N-gram	1-G	2-G	3-G	4-G	5-G	6-G
CIC2019	CSDPayload+CSPayload+N-gram	99.70	99.98	99.99	99.99	99.99	99.98	99.98
MIB2016	CSDPayload+CSPayload+N-gram	100.00	99.12	99.36	99.66	99.64	93.88	96.23
H2N-Payload	CSDPayload+CSPayload+N-gram	57.52	98.67	99.18	99.18	99.33	98.00	96.67
Average		85.74	99.26	99.51	99.61	99.65	97.29	97.63

3.6. Compare algorithm and result

Performance evaluation for all features in each dataset was also carried out in this study, with the results shown in Table 9. In addition, this study also tested the classification performance level for DDoS attack detection using combined features. The accuracy rate for the NN algorithm on CIC-2019 dataset are 99.99% respectively. The SVM algorithm achieved 99.84 and 99.54% for MIB-2016 and H2N-Payload dataset respectively. The KNN algorithm achieved 99.58% on H2N-Payload dataset.

Table 9. Performance evaluation for combining all features (hybrid features) using weight by correlation

Dataset	Number of features	Machine learning algorithms	Accuracy	Recall	Precision
CIC-2019	32	KNN	99.67	99.64	99.38
		NN	99.99	100.00	99.97
MIB-2016	17	SVM	99.84	99.60	100.0
		KNN	98.58	98.90	99.28
		NN	99.84	99.90	100.0
H2N-Payload	18	SVM	99.54	99.40	99.50
		KNN	99.58	98.90	99.99
		NN	99.44	99.40	99.30

4. CONCLUSION

The experimental results demonstrate that the SVM algorithm achieves the highest overall classification performance across the evaluated datasets. Specifically, SVM attained accuracy rates of 99.80% on CIC-2019, 99.74% on MIB-2016, and 99.64% on H2N-Payload, yielding an average accuracy of 99.73%. In comparison, the KNN algorithm achieved accuracies of 99.71, 91.66, and 94.06% on the same datasets, respectively, with an average of 95.14%. The NN model also performed strongly, with accuracies of 99.99, 99.64, and 99.33%, resulting in an average of 99.65%. Although the NN achieved the highest accuracy on the CIC-2019 dataset, SVM demonstrated the most consistent and highest average performance across all three datasets, making it the best-performing algorithm in this study. For future work, a more in-depth investigation using advanced deep learning architectures applied to the same datasets but with extended or alternative feature sets could further enhance DDoS detection capabilities and generalization.

ACKNOWLEDGMENTS

This work was supported by the Universiti Tun Hussein Onn Malaysia (UTHM) through Tier1 (votQ508) and also received support from industry.

FUNDING INFORMATION

This research was funded by Universiti Tun Hussein Onn Malaysia (UTHM) with grant number votQ508 and supported by practitioners from industry.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Andi Maslan	✓	✓	✓		✓	✓	✓	✓	✓		✓	✓	✓	
Cik Feresa Mohd Foozy	✓			✓		✓				✓		✓		✓
Kamaruddin Malik	✓			✓		✓				✓	✓	✓		
Mohamad														
Abdul Hamid		✓		✓	✓				✓	✓			✓	
Dedy Fitriawan						✓	✓			✓	✓			
Joni Hasugian	✓	✓	✓						✓				✓	

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

In this research there is no conflict of interest towards any party.

DATA AVAILABILITY

This research data is available on the official website of the University of New Brunswick (UNB) at <https://www.unb.ca/cic/datasets/ddos-2019.html>, which provides real-time data sets related to research in the field of network security.




REFERENCES

- [1] S. Sambangi, L. Gondi, and S. Aljawameh, "A feature similarity machine learning model for DDoS attack detection in modern network environments for industry 4.0," *Computers and Electrical Engineering*, vol. 100, 2022, doi: 10.1016/j.compeleceng.2022.107955.
- [2] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 1, 2019, doi: 10.1186/s42400-019-0038-7.
- [3] M. Aldwairi, W. Mardini, and A. Alhowaide, "Anomaly payload signature generation system based on efficient tokenization methodology," *International Journal on Communications Antenna and Propagation*, vol. 8, no. 5, pp. 421–429, 2018, doi: 10.15866/irecap.v8i5.12794.
- [4] I. Masud, K. Kusriani, and A. B. Prasetyo, "Distributed denial of service (DDoS) Attack Detection On Zigbee protocol using naive Bayes algorithm," *International Journal of Artificial Intelligence Research*, vol. 5, no. 2, pp. 157–167, 2021, doi: 10.29099/ijair.v5i2.214.
- [5] M. Zahid and T. S. Bharati, "Enhancing cybersecurity in IoT systems: a hybrid deep learning approach for real - time attack detection," *Discover Internet of Things*, vol. 5, no. 73, 2025, doi: 10.1007/s43926-025-00156-y
- [6] N. Bindra and M. Sood, "Evaluating the impact of feature selection methods on the performance of the machine learning models in detecting DDoS attacks," *Romanian Journal of Information Science and Technology*, vol. 23, no. 3, pp. 250–261, 2020.
- [7] A. Azhari, A. W. Muhammad, and C. F. M. Foozy, "Machine learning-based distributed denial of service attack detection on intrusion detection system regarding to feature selection," *International Journal of Artificial Intelligence Research*, vol. 4, no. 1, pp. 1–8, 2020, doi: 10.29099/ijair.v4i1.156.
- [8] J.-J. Kim, Y.-S. Lee, J.-Y. Moon, and J.-M. Park, "Network payload and correlation analysis in bigdata environments," *International Journal of Grid and Distributed Computing*, vol. 11, no. 3, pp. 109–124, 2018, doi: 10.14257/ijgdc.2018.11.3.10.
- [9] M. Aamir and S. M. A. Zaidi, "Clustering based semi-supervised machine learning for DDoS attack classification," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 4, pp. 436–446, 2021, doi: 10.1016/j.jksuci.2019.02.003.
- [10] M. Arshi, M. D. Nasreen, and K. Madhavi, "A survey of DDOS attacks using machine learning techniques," in *E3S Web of Conferences*, 2020, doi: 10.1051/e3sconf/202018401052.




- [11] F. S. D. L. Filho, F. A. F. Silveira, A. D. M. B. Junior, G. V.-Solar, and L. F. Silveira, "Smart detection: an online approach for DoS/DDoS attack detection using machine learning," *Security and Communication Networks*, vol. 2019, pp. 1–15, 2019, doi: 10.1155/2019/1574749.
- [12] M. Wang, Y. Lu, and J. Qin, "A dynamic MLP-based DDoS attack detection method using feature selection and feedback," *Computers and Security*, vol. 88, 2020, doi: 10.1016/j.cose.2019.101645.
- [13] A. Manna and M. Alkasassbeh, "Detecting network anomalies using machine learning and SNMP-MIB dataset with IP group," *2019 2nd International Conference on new Trends in Computing Sciences (ICTCS)*, Amman, Jordan, 2019, pp. 1-5, doi: 10.1109/ICTCS.2019.8923043.
- [14] I. Riadi, A. W. Muhammad, and Sunardi, "Network packet classification using neural network based on training function and hidden layer neuron number variation," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 6, pp. 1–4, 2017, doi: 10.14569/IJACSA.2017.080631.
- [15] K. Swapna and 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.
- [16] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Intrusion detection evaluation dataset (CIC-IDS2017)," *Canadian Institute for Cybersecurity*, 2017, Accessed: Jun. 17, 2018, [Online] Available: <https://www.unb.ca/cic/datasets/ids-2017.html>
- [17] 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.
- [18] M. Alkasassbeh, G. Al-Naymat, A. B.A, and M. Almseidin, "Detecting distributed denial of service attacks using data mining techniques," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 1, 2016, doi: 10.14569/IJACSA.2016.070159.
- [19] H. S. Obaid and E. H. Abeed, "DoS and DDoS attacks at OSI layers," *International Journal of Multidisciplinary Research and Publications*, vol. 2, no. 8, pp. 1–9, 2020, doi: 10.5281/zenodo.3610833.
- [20] H. Ren, S. Xiao, and H. Zhou, "A chi-square distance-based similarity measure of single-valued neutrosophic set and applications," *International Journal of Computers, Communications and Control*, vol. 14, no. 1, pp. 78–89, 2019, doi: 10.15837/ijccc.2019.1.3430.
- [21] A. Maslan, K. M. Mohamad, A. Hamid, H. Pangaribuan, and S. Sitohang, "Feature selection to enhance DDoS detection using hybrid n-gram heuristic techniques," *International Journal on Informatics Visualization*, vol. 7, no. 3, pp. 815-822, 2021, doi: 10.30630/ijov.7.3.1533.
- [22] 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.
- [23] K. Rahul and K. Heena, "Soft computing techniques for various image processing applications: a survey," *Journal of Electrical and Electronic Engineering*, vol. 8, no. 3, p. 71, 2020, doi: 10.11648/j.jee.20200803.11.
- [24] K. K. Verma, B. M. Singh, and A. Dixit, "A review of supervised and unsupervised machine learning techniques for suspicious behavior recognition in intelligent surveillance system," *International Journal of Information Technology (Singapore)*, vol. 14, no. 1, pp. 397–410, 2022, doi: 10.1007/s41870-019-00364-0.
- [25] S. Treadwell and Z. Mian, "A heuristic approach for detection of obfuscated malware," in *2009 IEEE International Conference on Intelligence and Security Informatics, ISI 2009*, 2009, pp. 291–299, doi: 10.1109/ISI.2009.5137328.
- [26] K. H. Z. and A. A. A. N. A. M. Alhammadi, "A review of the common DDoS Attack: types and protection approaches based on artificial intelligence," *Fusion: Practice and Applications*, no. December, pp. 8–14, 2021, doi: 10.54216/fpa.070101.
- [27] F. Khanum, P. S. Lakshmi, and H. V. Reddy, "Network analysis of multisource packet capture files using machine learning," in *2024 5th International Conference on Circuits, Control, Communication and Computing (I4C)*, 2024, pp. 119–124, doi: 10.1109/I4C62240.2024.10748490.

BIOGRAPHIES OF AUTHORS






Andi Maslan    received a degree in Informatics Engineering at the Budi Utomo Institute of Technology Jakarta (2004), a master's degree in computer science (Information Systems) at the STMIK Putera Batam (2011) and Ph.D. at Universiti Tun Hussein Onn Malaysia (UTHM) in the field of Information Technology focusing on network security in 2024. The author 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.






Cik Feresa Mohd Foozy    received the degree in Information Technology and Multimedia and the master's degree in Computer Science (Information Security) from Universiti Teknologi Malaysia (UTM), in 2006 and 2009, respectively, and the Ph.D. degree in Information Security from Universiti Teknikal Malaysia Melaka (UTeM), in 2017. She started her career as a Lecturer at the Department of Information Security and Web Technology, UTHM, in November 2011. She is currently an active researcher and has written and presented a number of papers in conferences and journals. She can be contacted at email: feresa@uthm.edu.my.






Kamaruddin Malik Mohamad    received the degree in Computer Science and master's degree in Computer Science (Information Security) from Universiti Teknologi Malaysia (UTM), in 1992 and 2003, the Ph.D., degree in Information Technology from Universiti Tun Hussein Onn Malaysia, in 2011. He started his career as a Lecturer at the Department of Information Security and Web Technology, UTHM, in 2004. His research interest includes file carving, steganography, secure data wiping, digital forensics triage, digital forensic analysis, metadata visualization and data redact. He can be contacted at email: malik@uthm.edu.my.






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.



Joni Hasugian    received a Bachelor of Computer Science graduate from Putera Batam University since 2015, has approximately 10 years of experience as a Network Engineer and Infrastructure, and has implementation knowledge in Cisco devices, Fortigate Firewall, Mikrotik, VMware, Windows Server, Software Installation, and Computer Hardware. As a holder of CCNA, CCNP, NSE4, MTCINE, and VCP-DCV certificates, and currently works at PT. Angkasa Pura Indonesia as a network engineer. He can be contacted at email: aeromic001@gmail.com.