

Performance Evaluation of Feature-Selection-Based Support Vector Machine Kernels for Real-Time DDoS Attack Detection via Orange3 Platform

Waleed Abdullah Haroun Miftah *

Department of Computer Science, Faculty of Science, University of Derna, Derna, Libya

* Corresponding author: Walidharoon79@gmail.com

تقييم أداء أنوية آلات ناقلات الدعم (SVM) القائم على اختيار الميزات للكشف عن هجمات DDoS في الوقت الفعلي عبر منصة Orange3

وليد عبد الله هارون مفتاح *

قسم الحاسوب، كلية العلوم، جامعة درنة، درنة، ليبيا

Received: 02-07-2026; Accepted: 14-09-2026; Published: 29-09-2026

Abstract

Distributed Denial-of-Service (DDoS) attacks severely threaten network stability. Early detection remains challenging due to massive traffic volumes that slow down algorithms and induce bias. This study evaluates Support Vector Machine (SVM) performance by comparing four kernels (Linear, RBF, Polynomial, and Sigmoid) and investigating feature filtering techniques to enhance classification efficiency. Employing a comparative experimental approach via the Orange3 visual programming environment, the methodology utilizes two standard datasets: CICDDoS2019 and CICDoS2017. These datasets were processed using balanced random sampling alongside two feature selection techniques: Information Gain (IG) and Information Gain Ratio (IGR). The results revealed sharp behavioral discrepancies between the datasets. For the CICDDoS2019 cohort, the IGR Sigmoid SVM model emerged as the top performer, achieving a Classification Accuracy (CA) of 90.1%, an F1-score of 89.2%, and excellent AUC stability at 90.8%. Conversely, hidden bias caused the RBF kernel's AUC curve to plunge to 1.8%. For the CICDoS2017 cohort, accuracy sharply declined, capping at 42.2% due to the algorithm's susceptibility to competitive noise from outdated data and crosstalk attacks. The study recommends integrating IGR as a mandatory preprocessing step for dimensionality reduction and cautions against relying solely on CA without AUC correlation. It advises avoiding traditional kernels like RBF in imbalanced environments and explores hybrid models for complex datasets like CICDoS2017.

Keywords: Cybersecurity, DDoS attacks, Support Vector Machine (SVM), Feature selection, Information Gain Ratio (IGR), Orange3 platform.

الملخص

تشكل هجمات حجب الخدمة الموزعة (DDoS) تهديداً خطيراً لاستقرار الشبكات، ولا يزال الكشف المبكر عنها يمثل تحدياً كبيراً بسبب الحجم الهائل لبيانات حركة المرور التي تبطئ الخوارزميات وتسبب التحيز. تهدف هذه الدراسة إلى تقييم أداء خوارزمية آلات ناقلات الدعم (SVM) عبر مقارنة أربعة أنوية (Linear و RBF و Polynomial و Sigmoid)، وبحث أثر تقنيات تصفية الميزات في تحسين كفاءة التصنيف. باتباع منهج تجريبي يقارن عبر بيئة البرمجة المرئية Orange3، اعتمدت الدراسة على مجموعتي بيانات قياسييتين: CICDDoS2019 و CICDoS2017، وتمت معالجتهما بواسطة المعاينة العشوائية المتوازنة وتقنيتين لاختيار الميزات: كسب المعلومات (IG) ونسبة كسب المعلومات (IGR). أظهرت النتائج تبايناً حاداً في السلوك بين مجموعتي البيانات؛ ففي مجموعة CICDDoS2019، تميز نموذج (IGR Sigmoid SVM) كأفضل نموذج محققاً دقة تصنيف (CA) بنسبة 90.1%، ومقياس F1 بنسبة 89.2%، واستقراراً ممتازاً للمساحة تحت المنحنى (AUC) عند 90.8%. وفي المقابل، أدى التحيز الخفي إلى هبوط حاد في منحنى AUC لنواة RBF إلى 1.8%. أما في مجموعة CICDoS2017، فقد انخفضت الدقة بشكل ملحوظ ولم تتجاوز 42.2% نظراً لحساسية الخوارزمية للضوضاء التنافسية الناتجة عن البيانات القديمة وهجمات التداخل. توصي الدراسة بدمج تقنية IGR كخطوة معالجة أولية إلزامية لتقليل الأبعاد، وتحذر من الاعتماد الكلي على مقياس CA دون ربطه بـ AUC، كما تنصح بتجنب الأنوية التقليدية مثل RBF في البيانات غير المتوازنة، واستكشاف النماذج الهجينة للتعامل مع البيانات المعقدة مثل CICDoS2017.

الكلمات المفتاحية: الأمن السيبراني، هجمات حجب الخدمة الموزعة (DDoS)، آلة ناقلات الدعم (SVM)، اختيار الميزات، نسبة كسب المعلومات (IGR)، منصة Orange3.

1. INTRODUCTION

The modern world is witnessing a rapid digital transformation that has led to the complete reliance of organizations and vital sectors on network infrastructure and cloud computing systems. With this widespread digital adoption, cyber threats have intensified and become more sophisticated, making information security and network stability a cornerstone for ensuring business continuity and maintaining national security for both institutions and countries.

Among these common threats, DDoS attacks stand out as one of the most destructive cyber weapons. They rely on overwhelming servers with massive amounts of fake traffic, rendering them inoperable and unavailable. Traditional, rule-based methods are no longer effective against these attacks due to their continuous evolution. The core problem lies in the fact that modern network traffic generates enormous amounts of data with extremely high dimensions and complexity. This presents a significant technical challenge: the slowness and low predictive accuracy of detection systems, leading to delayed response times and potentially devastating losses.

To address this challenge, scientific research has focused on employing artificial intelligence and machine learning techniques to build intelligent intrusion detection systems (IDS). The SVM algorithm stands out as a powerful classification solution; however, its efficiency depends critically on the type of kernel used and the quality of the input data. This study aims to bridge this research gap by providing a comparative applied evaluation of four different SVM kernels: Linear, RBF, Polynomial, and Sigmoid. It also examines the direct impact of feature selection and filtering techniques, such as IG and IGR, on reducing computational dimensions and accelerating processing time. This research is implemented experimentally using the Orange3 visual programming environment and two of the most realistic datasets: CICIDS2017 and CICDDoS2019, to arrive at the optimal model that achieves the highest accuracy and lowest response time in detecting attacks.

The scientific and practical value of this research lies in providing a comparative critical analysis that reveals the profound behavioral differences of the SVM algorithm with its various intentions when dealing with different generations and environments of DDoS attack data. The research offers original contributions that fill important research gaps in the fields of feature engineering and cybersecurity, and these are summarized in detail in the following points:

- Re-establishing the academic credibility of the Sigmoid kernel and proving the conditionality of its success: The research breaks the prevailing concept in previous literature that excludes the Sigmoid kernel and considers it ineffective in network classification; The applied study proved that combining this core with IGR technology resulted in a significant performance leap (in the CICDDoS2019 group), raising its accuracy from 32.8% to 90.1% and stabilizing the AUC curve at 90.8%, thus proving its superiority when features are properly filtered.
- Uncovering the RBF Core Bias Dilemma and the Hidden AUC Collapse: The research provided an important analytical finding that warns developers of intrusion detection systems against relying solely on the accuracy metric (CA). Experiments have shown that combining the RBF core with IGR technology gives a misleading impression of success due to recording a high classification accuracy (89.0%), while the AUC curve simultaneously collapses sharply to 1.8% as a result of the core falling into the trap of severe bias towards one class due to data imbalance.
- Identifying the effect of structural interference and noise between data generations (2017 vs. 2019): The research contributed to providing a scientific justification for the sharp difference in algorithm performance between the two groups; The study demonstrated that the SVM algorithm performs exceptionally well in the CICDDoS2019 environment due to the clear statistical signature of modern attacks, while its accuracy drops sharply below 42.2% in the CICDDoS2017 environment due to the heavy interference and noise generated by traditional flooding attacks. It also showed that filtering traditional features in overlapping environments can harm the model by deleting critical indicators.
- Furthermore, the study presented a visual, low-code pipeline: Orange3 demonstrated its effectiveness as a visual engineering and analysis tool capable of handling large and complex benchmark data files and accurately measuring the computational and time efficiency of four entredres. This enables cybersecurity analysts to quickly build and test models without getting bogged down in lengthy code complexities.

2. RELATED WORKS

Previous studies are extremely important, as they enable the researchers to review developments in their chosen topic. They can identify flaws or gaps upon which research can be built. Below are some of the most important recent studies:

Mohammed et al. [1] proposed a robust defense mechanism to secure IoT applications against DDoS threats. Utilizing a fraud detection dataset coupled with the K-means algorithm for feature selection, their empirical evaluation revealed that the SVM outperformed other models, achieving an outstanding classification accuracy of 99.85%. Similarly, Manjula and Mangla [2] addressed DDoS detection by developing a custom dataset and leveraging the distributed computing capabilities of Apache Spark. By applying classification algorithms via Spark's MLlib library, their framework successfully categorized network traffic, with the Random Forest (RF) classifier yielding the highest accuracy of 96.75%.

In another study, Rajendran et al. [3] explored machine learning applications for DDoS mitigation using the benchmark CICDDoS2019 dataset. Implemented on the Anaconda Jupyter platform, their work emphasized data preprocessing and visual analytics, concluding that the RF model achieved a peak accuracy of 99.10%. Focusing on Software-Defined Networking (SDN), Söğüt et al. [4] investigated DDoS detection within SDN-based SCADA systems. Utilizing a standard reference dataset, their findings demonstrated that the SVM classifier attained a strong accuracy rate of 97.2%.

Furthermore, Ikhsanudin et al. [5] introduced a hybrid approach for securing the MQTT protocol against DDoS attacks using a semi-supervised DBSCAN and SVM model. Evaluated across multiple datasets, including IoTID20 and CICDDoS2018, their model achieved a high detection accuracy of 99.6%. In a similar SDN context, Mohan et al. [6] simulated a network environment using Mininet and the Ryu controller to generate a real-time DDoS dataset. Their classification results indicated that the RF algorithm achieved an accuracy of 97.37% in distinguishing normal traffic from malicious patterns.

Recent advancements also highlighted the role of feature selection; Rui et al. [7] evaluated the performance of SVM and RF classifiers for DDoS detection in SDNs. By integrating Genetic Algorithm (GA) and Recursive Feature Elimination (RFE) for feature optimization, their results showed that RF reached an accuracy of 99%. Additionally, Rasheed and Rathore [8] benchmarked the performance of SVM using the KDD and KDD'99 datasets. By filtering the data down to six critical features, the SVM model achieved accuracies of 99.87% and 87.90% for the respective datasets. Patel et al. [9] conducted a comparative study between Binary Decision Trees (DT), XGBoost, and SVM using a dual-dataset approach for training and testing. Their outcomes indicated that while both XGBoost and Binary DT achieved an identical accuracy of 99%, the Binary DT exhibited superior computational efficiency by requiring significantly less execution time.

To address ML application constraints, Swati et al. [10] developed a custom dataset to evaluate machine learning models against DDoS assaults, where the DT algorithm demonstrated the highest classification performance with an accuracy of 96.51%. In the healthcare domain, Welteji et al. [11] focused on protecting medical environments from network disruptions using the CICIDS dataset. Their benchmark results revealed that the J48 decision tree model achieved a near-perfect accuracy of 99.98%. Similarly, Sabir [12] leveraged the CIC-DDoS2019 dataset by applying IG for feature selection, along with zero-instance attribute removal and instance randomization filters. The optimized framework showed that the J48 algorithm attained the highest accuracy of 98.31%.

Furthermore, Kannan et al. [13] evaluated network security using a custom-built dataset, demonstrating that DT, RF, and Extra Trees algorithms could reach an absolute accuracy of 100%. Raza et al. [14] proposed a feature-selection-based framework on the CICDDoS2019 dataset. By filtering zero-instance attributes and extracting the most relevant features via SelectKBest, univariate selection, and correlation analysis, their optimized RF model achieved a high detection accuracy of 99%. Finally, Hirsi et al. [15] introduced a novel ensemble learning framework combined with Principal Component Analysis (PCA) for optimized feature selection. Evaluated on both a custom-built dataset and the standard CIC-DDoS2019 dataset, their approach leveraged multiple integrated algorithms to enhance overall predictive robustness, resulting in a peak classification accuracy of 100% using the ENRF ensemble model.

In summary, the reviewed literature demonstrates the high capability of machine learning algorithms, ranging from traditional single classifiers to advanced ensemble models (e.g., Hirsi et al. [15]), in mitigating DDoS attacks across various environments such as IoT, SDN, and healthcare systems. However, a significant gap remains unaddressed: while several studies highlight SVM as a top-performing classifier (e.g., Mohammed et al., Söğüt et al., and Rasheed & Rathore [8]), they treat SVM as a single entity without investigating the structural impact of its kernel functions. Network traffic data during a DDoS attack is inherently complex and non-linear, meaning that

the choice of the kernel function (Linear, Polynomial, RBF, or Sigmoid) can critically dictate both the detection accuracy and the computational overhead. Furthermore, there is a lack of research utilizing interactive data mining environments like Orange3 to systematically benchmark these kernels under uniform validation constraints.

To address these gaps, this study provides a rigorous, empirical comparison of four distinct SVM kernel variants—Linear, Polynomial, RBF, and Sigmoid—implemented within the Orange3 framework. By partitioning the dataset using a 70% train and 30% test split approach, this research aims to evaluate these classifiers under a standardized validation setup to identify the most effective hyperplane configuration that optimizes DDoS detection performance while maintaining runtime efficiency.

3. MATERIALS & METHODS

In this section, the author will describe the used datasets and how to properly prepare it before the training and testing process. The author will also explain the algorithms targeted in this research. In addition, the author will explain the proposed algorithm, which will be used to evaluate the performance of it compare to traditional type of this algorithm.

3.1 Research framework

The model used in this research is summarized in following: First selects the dataset to be used, data preprocessing and feature engineering, and then applies an evaluation method (70% training and 30% testing). This results in identifying the best algorithm and the superior algorithms for each criterion. Figure (1) presents the model flow chart of the research methodology.

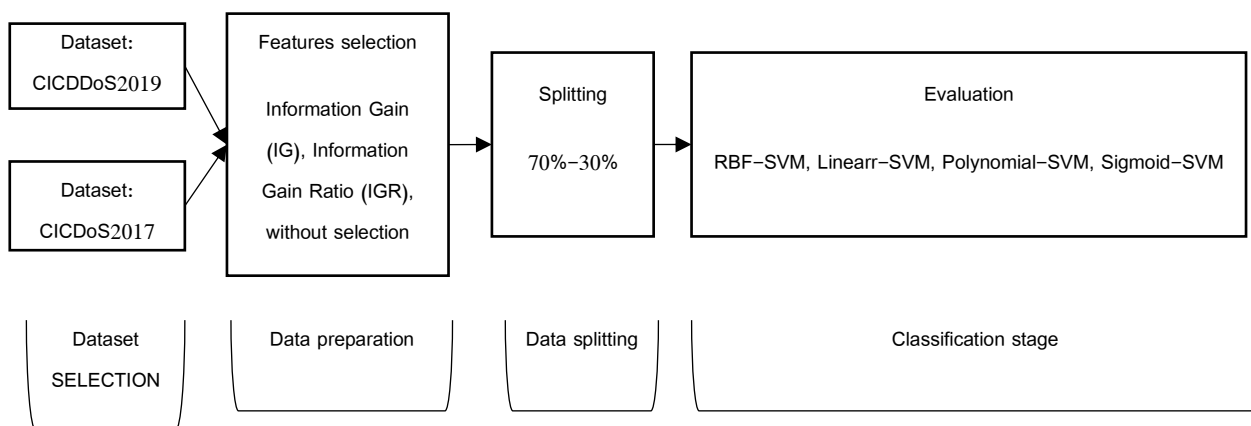


Figure 1. Model flow chart.

3.2 Datasets selection

This research employed the CICDDoS2019 dataset (79 features and 402208 instances) and CICDoS2017 dataset (78 features and 222914 instances), CICDDoS2019 dataset originally introduced by Sharafaldin et al. [16] at the Canadian Institute for Cybersecurity. The version used in this study was obtained from the publicly available repository on Kaggle [17], which provides a cleaned and reorganized edition of the dataset. Compared to the raw release, this version includes corrections for extreme class imbalance (reduced from ~4000:1 to ~9:1 attack-to-benign ratio), removal of duplicate and missing records, and conversion into optimized Parquet format for faster processing [17]. The CICDoS2017 dataset used in this study was obtained from the publicly available repository [18], which is a republished version of the original dataset created by Jazi et al. [19]. This study relies on Version 3 (V3) of the dataset, which includes preprocessing improvements such as data type corrections, removal of missing values and duplicates. The core structure of the dataset, including all 78 features and records, remains unchanged from the original release [18]. Since DoS and DDoS attacks share similar underlying patterns differing mainly in the number of sources involved successfully predicting a DoS attack can provide insights into detecting DDoS attacks as well. The datasets documentations do not disclose the preprocessing methods or the initial size of the raw data. Therefore, it is impossible to determine how records were filtered or reduced before publication. This presents a significant limitation when interpreting comparative results. The following figure illustrates the main types more clearly it from Canadian Institute for Cybersecurity [20]:

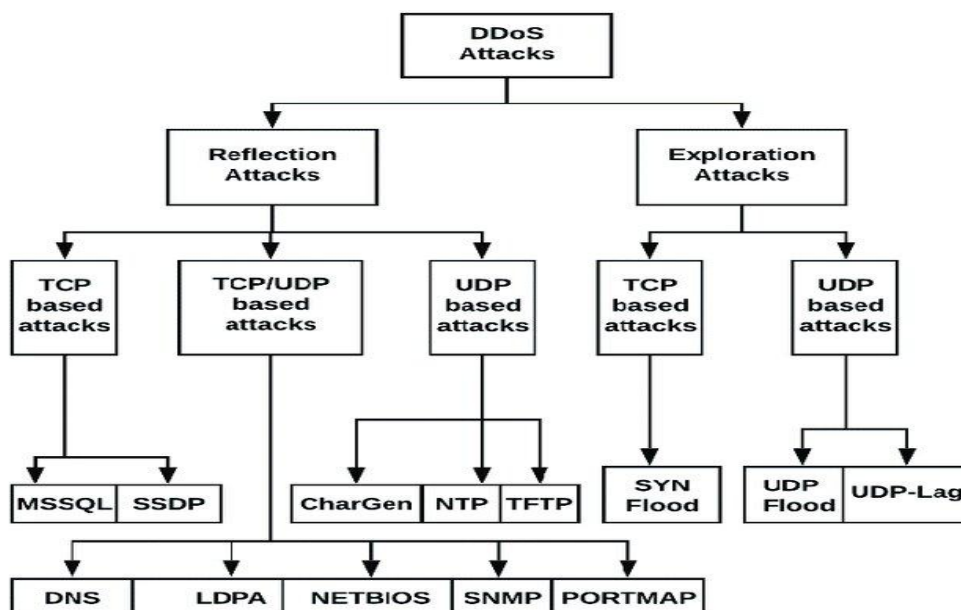


Figure 2. DDoS taxonomy in CICDDoS2019.

The types of DDoS attacks found in the CICDoS2017 dataset are shown in the following figure:

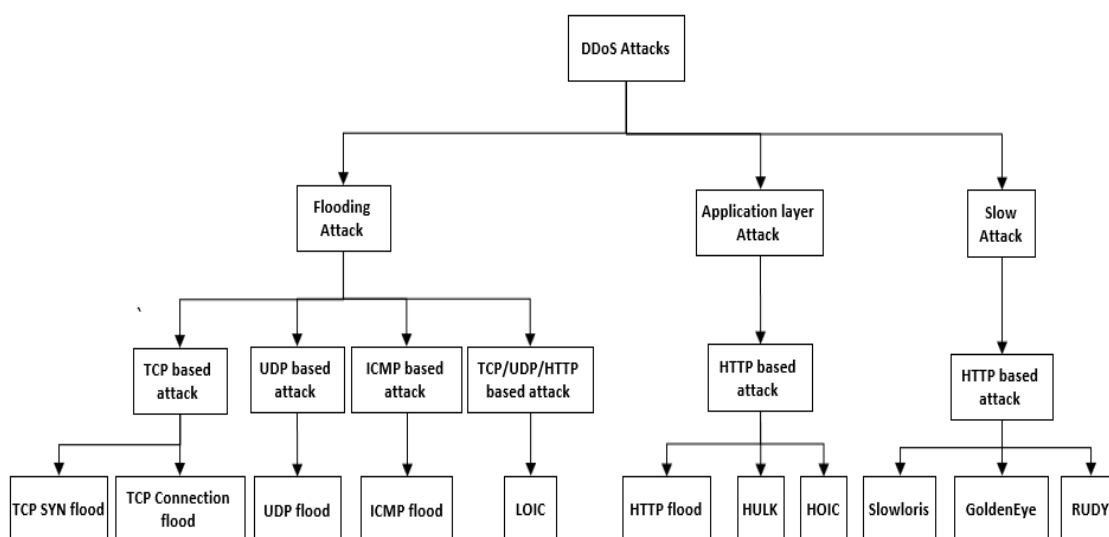


Figure 3. DDoS taxonomy in CICDoS2017.

Both datasets share the same columns in the same order, except for the Class, which exists only in CICDDoS2019. The remaining columns are common to both datasets, and they are showed in the table below:

Table 1. Columns of CICDDoS2019 and CICDoS2017 (features).

No.	Column Name
1	Protocol
2	Flow Duration
3	Total Fwd Packets
4	Total Backward Packets
5	Fwd Packets Length Total
6	Bwd Packets Length Total
7	Fwd Packet Length Max
8	Fwd Packet Length Min
9	Fwd Packet Length Mean
10	Fwd Packet Length Std
11	Bwd Packet Length Max

No.	Column Name
12	Bwd Packet Length Min
13	Bwd Packet Length Mean
14	Bwd Packet Length Std
15	Flow Bytes/s
16	Flow Packets/s
17	Flow IAT Mean
18	Flow IAT Std
19	Flow IAT Max
20	Flow IAT Min
21	Fwd IAT Total
22	Fwd IAT Mean
23	Fwd IAT Std
24	Fwd IAT Max
25	Fwd IAT Min
26	Bwd IAT Total
27	Bwd IAT Mean
28	Bwd IAT Std
29	Bwd IAT Max
30	Bwd IAT Min
31	Fwd PSH Flags
32	Bwd PSH Flags
33	Fwd URG Flags
34	Bwd URG Flags
35	Fwd Header Length
36	Bwd Header Length
37	Fwd Packets/s
38	Bwd Packets/s
39	Packet Length Min
40	Packet Length Max
41	Packet Length Mean
42	Packet Length Std
43	Packet Length Variance
44	FIN Flag Count
45	SYN Flag Count
46	RST Flag Count
47	PSH Flag Count
48	ACK Flag Count
49	URG Flag Count
50	CWE Flag Count
51	ECE Flag Count
52	Down/Up Ratio
53	Avg Packet Size
54	Avg Fwd Segment Size
55	Avg Bwd Segment Size
56	Fwd Avg Bytes/Bulk
57	Fwd Avg Packets/Bulk
58	Fwd Avg Bulk Rate
59	Bwd Avg Bytes/Bulk
60	Bwd Avg Packets/Bulk
61	Bwd Avg Bulk Rate
62	Subflow Fwd Packets
63	Subflow Fwd Bytes

No.	Column Name
64	Subflow Bwd Packets
65	Subflow Bwd Bytes
66	Init Fwd Win Bytes
67	Init Bwd Win Bytes
68	Fwd Act Data Packets
69	Fwd Seg Size Min
70	Active Mean
71	Active Std
72	Active Max
73	Active Min
74	Idle Mean
75	Idle Std
76	Idle Max
77	Idle Min
78	Label
79	Class (only in CICDDoS2019)

The CICDDoS2019 dataset contains two classification columns:

- The Label column, which specifies the detailed attack type (e.g., UDP or specific category).
- The Class column, which distinguishes between Benign and DDoS.

Since this study does not focus on identifying the exact type of attack, the Class column was excluded, and the values of the Label column in both datasets were reformulated into two categories:

- Benign.
- DDoS.

This adjustment ensured full consistency in column naming and structure between both datasets (CICDDoS2019 and CICDoS2017), thereby facilitating fair comparison and simplifying the prediction task by focusing on the fundamental distinction between normal traffic and malicious activity. Furthermore, reframing the problem as a binary classification task. This contributes to a clearer analysis and more reliable comparison of algorithmic performance.

3.3 Data preparation

This section covers the initial data preparation that will be applied during the training and testing phases of the algorithms. In some experiments, no initial data preparation will be applied to the data as a baseline

- Feature Selection

Information Gain (IG) is used to evaluate the ability of each feature to classify data based on reducing the degree of randomness, or what is known as entropy. Entropy is calculated before data partitioning using the following equation:

$$E = - \sum_{i=1}^c p_i \log_2 p_i$$

Where 'C' represents the number of classes in the target column, and 'p_i' is the probability (or percentage) of class 'i' appearing in the data. Based on this, the information gain of any feature is calculated by subtracting the weighted entropy after partitioning from the total entropy before partitioning, according to the equation:

$$\text{Information Gain} = \text{Entropy before split} - \text{Weighted entropy after split}$$

The closer the information gain value is to 1, the more effectively the feature can separate and classify the data, while lower values indicate higher dispersion and randomness.

Information Gain Ratio (IGR): Although effective in "information gain", it has the disadvantage of bias towards features that contain a large number of unique values (such as ID numbers or dates). To address this bias, the Information Gain Ratio (IGR) is used, which normalizes the information gain value by dividing it by a metric called "split information":

$$\text{Information Gain Ratio} = \frac{\text{Information Gain}}{\text{Split Information}}$$

Split information, which measures the variability and dispersion of values for the same feature, is calculated using the following formula:

$$Split\ Information = - \sum_{i=1}^u \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

Where $|S|$ represents the total number of records, and $|S_i|$ represents the number of records that possess the unique value. (i) represents the value of the evaluated feature, while (U) represents the total number of unique values for that feature. This ratio helps in selecting features that provide better generalization when applying the model to new data.

The selected features are presented in the table below:

Table 2. Selected features of the experiment.

Feature selection method	No of features	Final features selected
Without feature selected	78	All features
Information Gain with CICDDoS2019 dataset	15	Bwd Packets/s, Init Fwd Win Bytes, Down/Up Ratio, Bwd Packet Length Max, Avg Bwd Segment Size, Bwd Packet Length Mean, Subflow Bwd Packets, Total Backward Packets, Bwd IAT Mean, Bwd IAT Max, Bwd IAT Total, Bwd Header Length, Bwd IAT Min, Avg Fwd Segment Size, Fwd Packet Length Mean
Information Gain Ratio with CICDDoS2019 dataset	15	URG Flag Count, Down/Up Ratio, CWE Flag Count, RST Flag Count, Fwd PSH Flags, Bwd Packets/s, Init Fwd Win Bytes, Bwd Packet Length Std, Bwd Packet Length Max, Avg Bwd Segment Size, Bwd Packet Length Mean, Subflow Bwd Packets, Total Backward Packets, Bwd Header Length, Init Bwd Win Bytes
Information Gain with CICDoS2017 dataset	15	Idle Max, Idle Mean, Idle Min, Fwd Packets/s, Active Min, Active Max, Bwd Packets/s, Active Mean, Flow IAT Max, Flow Packets/s, Flow IAT Mean, Flow Duration, Idle Std, Flow IAT Std, Fwd IAT Std
Information Gain Ratio with CICDoS2017 dataset	15	Idle Std, Active Min, Active Max, Active Mean, Active Std, Idle Max, Idle Mean, Idle Min, Fwd Packets/s, Bwd Packets/s, SYN Flag Count, Flow IAT Max, Flow Packets/s, Flow IAT Mean, Flow Duration

3.4 Data Splitting

To ensure a genuine and impartial evaluation of the models and prevent overfitting:

- Stratified Split: The data was split (70% for training and 30% for independent testing).
- Maintaining Distribution: Stratification was used to ensure that the training and testing samples retained the same proportions and distribution of the attack and non-attack categories as in the original database.

3.5 Evaluation Stage & Classifiers

In this stage, the predictive performance of the SVM algorithm is evaluated and compared using four different kernel functions to detect DDoS attacks. The implementation and evaluation process are carried out within the Orange3 data mining framework. SVM is selected due to its high efficiency in high-dimensional spaces and its capability to find an optimal hyperplane that separates normal traffic from malicious DDoS traffic.

To handle the non-linear nature of network traffic data, the following four SVM kernel variants are deployed and evaluated:

- **Linear SVM:** This kernel is used when the data is linearly separable. It separates the classes using a straight line or a flat hyperplane, offering low computational complexity.
- **Polynomial SVM:** This kernel represents a non-linear classifier by computing the relationships between features up to a specified degree, which helps in capturing complex interactions between network traffic attributes.

- **Radial Basis Function (RBF) SVM:** As one of the most widely used kernels, RBF maps the network data into an infinite-dimensional space. It is highly effective in handling highly non-linear decision boundaries between normal and attack patterns.
- **Sigmoid SVM:** Derived from neural network concepts, this kernel acts similarly to a two-layer perceptron model, utilizing a hyperbolic tangent function to construct the decision boundary.

3.6 Experiments setup

This research used an HP Probook 650 G2 laptop running Windows 10 64-bit. The processor is an Intel(R) Core(TM) i5-6300U CPU @ 2.40 GHz. The installed RAM is 16 GB. Orange3 was used to run the experiments using the CICDoS2017 and CICDDoS2019 datasets.

4. RESULTS

The experimental results are presented in the following table and the two figures below it, followed by a comprehensive analytical review in the discussion section:

Table 3. Results of the experiment.

Exp No	Dataset	Feature Selection Algorithm	Model	Train	Test	AUC	CA	F1	Prec	Recall
3	CICDDoS2019	Information gain ratio	Sigmoid-SVM	41.149	2.456	90.8%	90.1%	89.2%	91.2%	90.1%
3	CICDDoS2019	Information gain ratio	RBF-SVM	46.11	7.756	1.8%	89.0%	87.8%	90.1%	89.0%
3	CICDDoS2019	Information gain ratio	Lineal-SVM	25.67	1.544	4.1%	80.4%	74.9%	84.2%	80.4%
1	CICDDoS2019	Without feature selection	Lineal-SVM	45.755	3.371	57.1%	77.8%	70.0%	81.9%	77.8%
2	CICDDoS2019	Information gain	RBF-SVM	45.558	7.325	54.3%	77.3%	74.4%	74.4%	77.3%
2	CICDDoS2019	Information gain	Lineal-SVM	26.608	1.464	75.7%	75.9%	65.8%	81.5%	75.9%
3	CICDDoS2019	Information gain ratio	Polynomial-SVM	33.349	1.424	3.4%	75.8%	65.4%	81.3%	75.8%
1	CICDDoS2019	Without feature selection	Polynomial-SVM	52.3	2.673	3.3%	75.7%	65.2%	79.7%	75.7%
2	CICDDoS2019	Information gain	Polynomial-SVM	31.4	1.873	12.6%	75.7%	65.2%	79.9%	75.7%
2	CICDDoS2019	Information gain	Sigmoid-SVM	42.838	2.914	52.6%	75.1%	67.0%	67.5%	75.1%
1	CICDDoS2019	Without feature selection	RBF-SVM	72.113	8.609	75.2%	68.9%	69.3%	69.9%	68.9%
4	CICDoS2017	Without feature selection	RBF-SVM	85.582	25.378	38.8%	42.2%	54.5%	86.5%	42.2%
4	CICDoS2017	Without feature selection	Sigmoid-SVM	33.896	2.409	66.8%	39.3%	51.7%	85.8%	39.3%
1	CICDDoS2019	Without feature selection	Sigmoid-SVM	65.54	4.55	37.3%	32.8%	36.2%	50.3%	32.8%
5	CICDoS2017	Information gain	Sigmoid-SVM	22.035	1.764	70.1%	18.7%	22.8%	93.6%	18.7%
6	CICDoS2017	Information gain ratio	Sigmoid-SVM	22	1.452	59.1%	18.5%	24.8%	82.1%	18.5%
5	CICDoS2017	Information gain	RBF-SVM	24.181	3.852	22.4%	16.6%	23.0%	76.9%	16.6%
6	CICDoS2017	Information gain ratio	RBF-SVM	23.175	3.997	17.3%	16.1%	23.1%	73.9%	16.1%
4	CICDoS2017	Without feature selection	Lineal-SVM	25.251	1.034	44.6%	10.4%	9.4%	83.9%	10.4%
4	CICDoS2017	Without feature selection	Polynomial-SVM	26.84	1.381	47.5%	9.1%	6.3%	93.1%	9.1%

Exp No	Dataset	Feature Selection Algorithm	Model	Train	Test	AUC	CA	F1	Prec	Recall
6	CICDoS2017	Information gain ratio	Polynomial-SVM	12.689	0.846	53.3%	8.0%	4.1%	90.8%	8.0%
5	CICDoS2017	Information gain	Lineal-SVM	12.347	0.825	18.3%	7.6%	3.2%	92.6%	7.6%
6	CICDoS2017	Information gain ratio	Lineal-SVM	12.061	0.72	16.4%	7.4%	5.3%	61.0%	7.4%
5	CICDoS2017	Information gain	Polynomial-SVM	15.72	0.891	19.8%	6.5%	1.1%	89.1%	6.5%

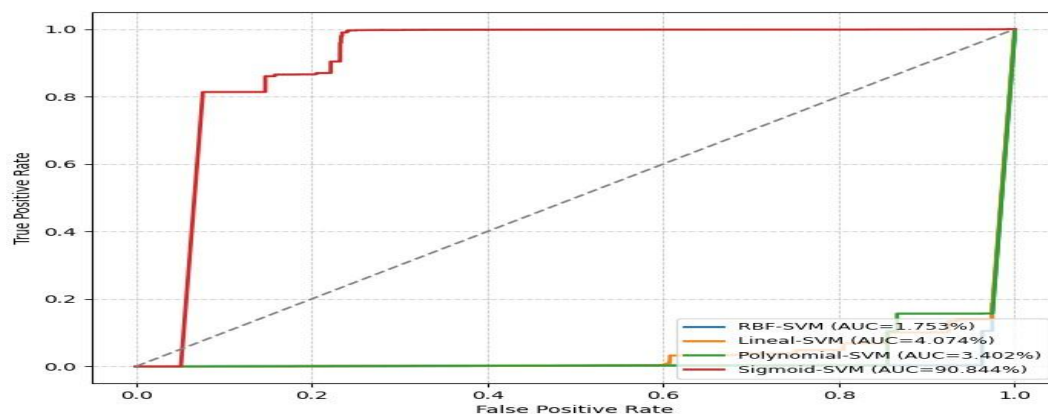


Figure 4. ROC curve of the experiment.

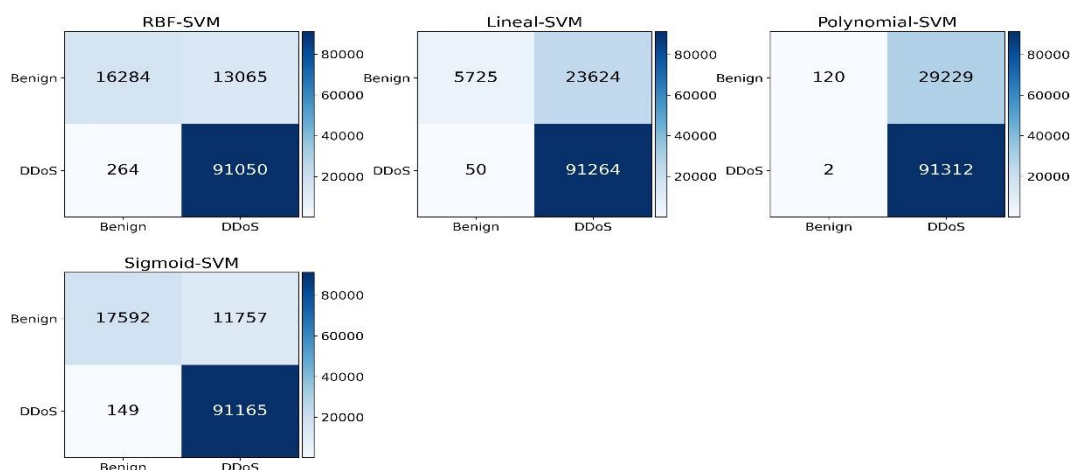


Figure 5. Confusion matrix of the experiment.

5. DISCUSSION

This section provides a critical analysis and scientific interpretation of the experimental results obtained from 6 experiments, which included the evaluation of four kernels of the SVM algorithm and a comparison of the effect of feature filtering techniques (IG and IGR) on the two benchmark datasets, CICDDoS2019 and CICDoS2017.

5.1 Comparative Analysis of the Datasets (CICDDoS2019 vs. CICDoS2017 Paradox):

The results show a sharp and fundamental behavioral divergence in the performance of the classification algorithms between the two datasets: In the CICDDoS2019 dataset: the models achieved high classification accuracy (CA) and excellent balance across performance metrics, peaking at 90.1% with the IGR Sigmoid-SVM model. In the CICDoS2017 dataset: there was a sharp and widespread collapse in accuracy, with the best result not exceeding 42.2% (using RBF without feature filtering), and accuracy deteriorating to below 10% with the remaining kernels. When applying feature filtering, the scientific and cyber justification for this discrepancy is as follows: This fundamental difference stems from the topological structure and nature of the attacks in both groups: Network fingerprint clarity in 2019: The CICDDoS2019 group focuses on modern application-layer DDoS attacks and reflection attacks (such as NTP, DNS, and LDAP). These attacks are characterized by very sharp and clear

statistical features in the packets, making it easier for the SVM algorithm to establish clear decision boundaries between legitimate and malicious traffic. Network noise and interference in 2017: In contrast, the CICDoS2017 group includes traditional and complex flooding attacks such as Slowloris, Heartbleed, and SYN Flood. These attacks rely on exploiting protocols and slowly and stealthily mimicking the behavior of legitimate users, resulting in massive statistical interference and high noise between the two categories (beneficial vs. DDoS). Accordingly, the mathematical separation boundaries of the SVM algorithm failed to distinguish between them, and when feature filtering techniques (IG/IGR) were applied, they deleted critical and small features that contributed to detection, leading to a complete collapse of computational accuracy in 2017.

5.2. Feature Selection Impact:

Experiments proved that feature engineering does not always lead to an absolute performance improvement, but its effect depends on its compatibility with the kernel used: Sigmoid Kernel and the Breakthrough: In the 2019 group, the IGR technique caused a qualitative leap for the Sigmoid kernel, as its accuracy increased from 32.8% (without feature filtering) to 90.1%. This is mathematically explained by the fact that IGR successfully eliminated redundant and variable features with high bias, allowing the sigmoid function to plot an ideal and stable classification curve. Performance deteriorated with complex clusters: In stark contrast, in the 2017 cluster, the application of IG and IGR reduced the classification accuracy of the Linear kernel from 10.4% to 7.4%, and the Polynomial kernel from 9.1% to 6.5%. This demonstrates the superiority of full features in the case of nested attacks, where the algorithm needs all available indicators to attempt separation.

5.3 AUC Curve Paradox and Model Bias (AUC Collapse):

The results reveal a highly significant statistical phenomenon that warns against relying on the abstract accuracy (CA) metric: In Experiment No. 3 of the CICDDoS2019 cohort, the RBF-SVM model with IGR achieved a high classification accuracy of 89.0% and an F1 metric of 87.8%. However, the AUC curve collapsed sharply to 1.8%. Technical Interpretation: This stark discrepancy indicates that the kernel has fallen into the trap of majority class bias due to imbalanced data. The model began classifying all traffic as normal (or vice versa) to achieve high abstract accuracy, while completely failing to sort probabilities and actually distinguish between the two categories, as revealed by the very low accuracy of the AUC curve. In contrast, the Sigmoid kernel with IGR demonstrated complete stability, as its high accuracy (90.1%) was accompanied by a distinct and stable AUC curve of (90.8%), making it the winning and only reliable model for use in real-world security environments.

5.4 Computational Complexity:

Regarding training and testing time, the results showed the following: The Polynomial and Linear kernels were consistently the fastest in all trials (with Polynomial training time with feature filtering ranging from 12 to 33 seconds), due to the simplicity of the mathematical operations for the separation lines. The RBF kernel was the most time-consuming and resource-intensive in both groups (recording the highest training time of 85.58 seconds and a testing time of 25.37 seconds in the group). (2017), due to the complexity of dot product calculations in the infinite spaces created by this kernel, making it impractical for systems that require real-time detection.

6. CONCLUSION

This research presents a comparative and comprehensive evaluation of the efficiency of SVM algorithm with its four intents (Linear, RBF, Polynomial, and Sigmoid) when combined with feature filtering techniques (IG and IGR), with the aim of improving the detection of DDoS attacks. The 6 experiments were conducted using the Orange3 visual programming environment and the two benchmark datasets CICDDoS2019 and CICDoS2017 to ensure a diversity of attack types and data structures. The results showed significant variation, reflecting the impact of network data types on algorithm behavior. The IGR Sigmoid-SVM model achieved exceptional success in the CICDDoS2019 dataset, achieving a classification accuracy of 90.1%, an F1 score of 89.2%, and a remarkably stable AUC curve of 90.8%, outperforming the traditional intents. In contrast, the research revealed a major technical challenge in the CICDoS2017 dataset; the accuracy collapsed sharply and did not exceed 42.2% as a result of high noise and severe interference between normal traffic and complex attacks. The research also proved the danger of relying on the abstract accuracy (CA) metric after observing the collapse of the RBF core AUC curve to 1.8% as a result of hidden statistical bias.

7. LIMITATION

Despite the strong and detailed results achieved by the research, there are several fundamental limitations that restrict the possibility of generalizing these results absolutely to all live intrusion detection systems, and these are summarized in the following points:

- The nature and structure of the datasets (Dataset-Specific Characteristics): The results are derived exclusively from the CICDDoS2019 and CICDoS2017 datasets. These datasets, although accurate, were generated in simulated sandbox environments. Traffic in this environment differs from that of real-world enterprise networks, which have more complex and dynamic user behavior patterns and protocol diversity, potentially impacting the consistency and accuracy of the winning algorithm.
- Lack of Model Cross-Adaptability: Experiments revealed significant performance variations; the model that achieved exceptional success with 2019 data (Sigmoid + IGR with 90.1% accuracy) collapsed completely with 2017 data (18.5%). This inconsistency demonstrates that the model is overly sensitive to the temporal and structural patterns of the attack, meaning it cannot be generalized as a one-size-fits-all solution without retraining and pre-configuring each network individually.
- Visual Development Environment (Platform and Low-Code Limitations): While relying on the Orange3 environment provided high speed and efficiency in visual design, it imposed limitations on the precise allocation of hyperparameters compared to writing raw code using libraries like Scikit-Learn in Python. Furthermore, models designed in these environments do not allow for immediate production deployment to routers or firewalls for real-time wire-speed filtering.

8. FUTURE WORK

Based on the findings and challenges revealed by this research, the following research directions are recommended for the future development of intrusion detection systems:

- Developing hybrid classification models: It is recommended to integrate the SVM algorithm with deep learning algorithms (such as LSTM or CNN) to handle complex and nested clusters like CICDoS2017, leveraging deep learning's ability to extract hidden features in high-noise environments. Advanced data imbalance mitigation: In future research, we suggest integrating advanced resampling techniques such as SMOTE or ADASYN before the classification stage with the RBF core to protect it from hidden bias and probability ranking failure (AUC collapse).
- Exploring adaptive feature selection techniques: The study recommends testing intelligent optimization-based feature selection algorithms (such as genetic algorithms or particle swarm optimization) and comparing them with IG and IGR to determine their ability to retain small and critical features in nested networks.
- Testing in environments Real-Time Deployment: It is proposed to transfer the winning model and develop it within a software environment that allows real-time testing on live network traffic to measure the resilience of the temporal and computational efficiency of intents against real-time attacks.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

REFERENCES

1. B. H. Mohammed, H. Sallehudin, N. Safie, and M. Satar, "Anomaly detection of distributed denial of service (DDoS) in IoT network using machine learning," Research Square, Preprint, Nov. 14, 2023. doi: 10.21203/rs.3.rs-3496063/v1.
2. H. T. Manjula and N. Mangla, "An approach to on-stream DDoS blitz detection using machine learning algorithms," Materials Today: Proceedings, vol. 47, pp. 4811-4815, 2021, doi: 10.1016/j.matpr.2021.05.517.
3. T. Rajendran, E. Abishekraj, and U. Dhanush, "Improved intrusion detection system that uses machine learning techniques to proactively defend ddos attack," in ITM Web of Conferences, vol. 56, 2023, Art. no. 05011, doi: 10.1051/itmconf/20235605011.
4. E. Söğüt, A. Tekerek, and O. A. Erdem, "Machine learning-based DDoS attack detection on SDN-based SCADA systems," Gazi Journal of Engineering Sciences, vol. 9, no. 3, pp. 596-611, 2024, doi: 10.30855/gjes.2024.09.03.023.

5. M. Ikhsanudin, V. Suryani, and R. R. Pahlevi, "DDoS Attack Detection on MQTT Protocol Using Semi-supervised DBSCAN and Support Vector Machine Model," in 2023 11th International Conference on Information and Communication Technology (ICoICT), Melaka, Malaysia, 2023, pp. 301-306, doi: 10.1109/ICoICT58202.2023.10262594.
6. V. Mohan, B. K. Madhavi, and S. B. Kishor, "Detection of UDP SYN Flood DDoS Attack Using Random Forest Machine Learning Algorithm in a Simulated Software Defined Network," in ICT for Intelligent Systems (Smart Innovation, Systems and Technologies, vol. 361), J. Choudrie, P. N. Mahalle, T. Perumal, and A. Joshi, Eds. Singapore: Springer, 2023, pp. 455-464, doi: 10.1007/978-981-99-3982-4_44.
7. H. Z. Rui, T. Y. Chien, L. X. Ee, C. W. San, and L. T. Yi, "Comparison of the use of support vector machine (svm) & random forest algorithms (rf) for ddos attack detection," International Journal of Research and Innovation in Social Science, vol. 9, no. 1, pp. 1126-1138, Jan. 2025.
8. S. Rasheed and M. S. Rathore, "Support vector machine based ddos detection and mitigation in software defined networks," Journal of Innovative Computing and Emerging Technologies, vol. 4, no. 2, 2024.
9. M. Patel, P. P. Amritha, V. B. Sudheer, and M. Sethumadhavan, "DDoS attack detection model using machine learning algorithm in next generation firewall," Procedia Computer Science, vol. 233, pp. 175-183, 2024, doi: 10.1016/j.procs.2024.03.207.
10. J. Swati, P. Nitin, S. Singh, A. Sinha, V. Sirvi, and S. Srivastava, "DDoS Attack Detection Using Machine Learning," in Decision Intelligence (Lecture Notes in Electrical Engineering, vol. 1079), B. K. Murthy, B. V. R. Reddy, N. Hasteer, and J.-P. Van Belle, Eds. Singapore: Springer, 2023, pp. 385-394, doi: 10.1007/978-981-99-5997-6_34.
11. B. T. Welteji, B. Tiwari, S. D. Kebede, S. Gupta, and V. Tiwari, "DDoS Attack Detection Using Predictive Machine Learning (ML) Algorithms in Wireless Body Area Network Environments," in IoT in Healthcare Systems, 1st ed., B. Tiwari, Ed. Boca Raton, FL, USA: CRC Press, 2023, pp. 249-274, doi: 10.1201/9781003145035-12.
12. M. Y. Sabir, "DDoS Attacks Detection using Machine Learning," M.Eng. report, Dept. Elect. Comput. Eng., Univ. Victoria, Victoria, BC, Canada, 2023.
13. C. Kannan, R. Muthusamy, V. Srinivasan, V. Chidambaram, and K. Karunakaran, "Machine learning based detection of DDoS attacks in software defined network," Indonesian Journal of Electrical Engineering and Computer Science, vol. 32, no. 3, pp. 1503-1511, Dec. 2023, doi: 10.11591/ijeecs.v32.i3.pp1503-1511.
14. M. S. Raza, M. N. A. Sheikh, I. S. Hwang, and M. S. Ab-Rahman, "Feature-selection-based DDoS attack detection using AI algorithms," Telecom, vol. 5, no. 2, pp. 333-346, Apr. 2024, doi: 10.3390/telecom5020016.
15. A. Hirsi, L. Audah, A. Salh, N. M. Sahar, M. A. Alhartomi, and S. Ahmed, "Detecting low-rate ddos attacks in sdn using ensemble machine learning techniques," in 2024 IEEE 22nd Student Conference on Research and Development (SCOREd), Dec. 2024, pp. 299-304, doi: 10.1109/SCOREd62283.2024.10853509.
16. Sharafaldin, I., Lashkari, A. H., Hakak, S., & Ghorbani, A. A. (2019, October). Developing realistic distributed denial of service (DDoS) attack dataset and taxonomy. In 2019 International Carnahan Conference on Security Technology (ICCST) (pp. 1–8). IEEE.
17. D'hooge, L. (2022a). CICDDoS2019 dataset [Data set]. Kaggle. Retrieved September 1, 2025, from <https://www.kaggle.com/datasets/dhoogla/cicddos2019>
18. D'hooge, L. (2022b). CICDoS2017 dataset [Data set]. Kaggle. Retrieved September 1, 2025, from <https://www.kaggle.com/datasets/dhoogla/cicdos2017>
19. Jazi, H. H., Gonzalez, H., Stakhanova, N., & Ghorbani, A. A. (2017). Detecting HTTP-based application layer DoS attacks on web servers in the presence of sampling. Computer Networks, 121, 25–36.
20. Canadian Institute for Cybersecurity. (n.d.). DDoS evaluation dataset (CIC-DDoS2019). University of New Brunswick. Retrieved August 22, 2025, from <https://www.unb.ca/cic/datasets/ddos-2019.html>.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of LJCAS and/or the editor(s). LJCAS and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.