Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.038003
Article

A Machine Learning-Based Distributed Denial of Service Detection Approach for Early Warning in Internet Exchange Points

Salem Alhayani* **and Diane R. Murphy**

School of Technology and Innovation College of Business, Innovation, Leadership, and Technology (BILT), Marymount University, Arlington, Virginia, 22207, USA

*Corresponding Author: Salem Alhayani. Email: Alhayani@gmail.com

Received: 24 November 2022; Accepted: 24 May 2023; Published: 30 August 2023

Abstract: The Internet service provider (ISP) is the heart of any country's Internet infrastructure and plays an important role in connecting to the World Wide Web. Internet exchange point (IXP) allows the interconnection of two or more separate network infrastructures. All Internet traffic entering a country should pass through its IXP. Thus, it is an ideal location for performing malicious traffic analysis. Distributed denial of service (DDoS) attacks are becoming a more serious daily threat. Malicious actors in DDoS attacks control numerous infected machines known as botnets. Botnets are used to send numerous fake requests to overwhelm the resources of victims and make them unavailable for some periods. To date, such attacks present a major devastating security threat on the Internet. This paper proposes an effective and efficient machine learning (ML)-based DDoS detection approach for the early warning and protection of the Saudi Arabia Internet exchange point (SAIXP) platform. The effectiveness and efficiency of the proposed approach are verified by selecting an accurate ML method with a small number of input features. A chi-square method is used for feature selection because it is easier to compute than other methods, and it does not require any assumption about feature distribution values. Several ML methods are assessed using holdout and 10-fold tests on a public large-size dataset. The experiments showed that the performance of the decision tree (DT) classifier achieved a high accuracy result (99.98%) with a small number of features (10 features). The experimental results confirm the applicability of using DT and chi-square for DDoS detection and early warning in SAIXP.

Keywords: Internet exchange point; Saudi Arabia IXP (SAIXP); distributed denial of service; chi-square; feature selection; machine learning

1 Introduction

An Internet service provider (ISP) is an entity that offers Internet services in any country. It plays a significant role in Internet infrastructure. The Internet exchange point (IXP) is an infrastructure that exchanges Internet traffic between ISPs and content delivery networks (CDNs). Internet

threats/attacks are evolving daily with new techniques [1]. One such attack is the distributed denial of service (DDoS) [2], which presents a serious threat to the Internet. Such attacks continue to increase in frequency, size, and complexity. Malicious actors in DDoS attacks control many infected machines known as botnets. Botnets are used to send numerous fake requests to overwhelm the resources of victims. Consequently, these resources are unavailable to legitimate users. Therefore, such attacks present a major devastating security threat on the Internet to date. Developing a platform to detect DDoS attacks is an important and challenging task [3].

This study proposes a platform to detect DDoS attacks at the Saudi Arabia Internet exchange point (SAIXP). DDoS detection and mitigation within SAIXP occur with many challenges. First, the IXP handles huge terabytes of traffic per second [4]. This requires an optimized, efficient, and effective detection mechanism. Second, the IXP is considered a neutral environment. The IXP must be careful when applying the mitigation of DDoS attacks and filtering traffic. For example, there is a question as to whether the IXP implementation mechanism should detect malicious Internet protocol (IP) spoofing and/or not look at any source-based filtering. Third, because the IXP network is a complex infrastructure with many surface attacks from the DDoS perspective, it is advantageous to use the DDoS early detection platform as an add-on in such infrastructure. Kaspersky Lab conducted a survey study on more than 5,200 business professionals [5], and the following results were obtained.

- On average, DDoS attacks cost companies $2 million, while they cost small- and medium-sized businesses (SMBs) $120,000.
- Compared to $1.6 million spent as a cost from companies for responding to DDoS attacks in 2016, the cost in 2017 increased to $2.3 million. However, the financial cost implications of responding to DDoS attacks in 2016 and 2017 were $106,000 and $123,000, respectively.
- Moreover, for all responders, the outcome is reported in the following points:
  - 33% identified the biggest financial strain caused by DDoS attacks as the expense of defending and restoring services.
  - 25% cited the cost of purchasing a backup or offline system as the main burden when online services are down.
  - 23% of respondents claimed that DDoS attacks directly resulted in a loss of revenue and business potential.
  - 22% of respondents identified the loss of reputation with customers and business partners as a direct result of DDoS attacks.

Traditionally, there are three machine learning (ML) approaches: supervised, semi-supervised, and unsupervised [6]. The detection method in a supervised manner requires a training dataset to discover anomalies. Both input variables and result classes are included in the training dataset. The hidden functions are extracted from the trained dataset, and the class of input incoming traffic variables is predicted. Linear regression and classification techniques are the most common supervised learning methods. Linear regression is a supervised method of learning that is typically used to estimate, forecast, and classify quantitative data relations. Classification techniques focus on predicting a qualitative response through data analysis and the recognition of patterns. The most commonly used classification algorithms are k-nearest neighbors (KNN), decision trees (DTs), naïve Bayes (NB), and support vector machines (SVM).

In the unsupervised type, the detection method can learn hidden functions from a given unlabeled dataset and identify the pattern of anomalies, but it yields less accurate detection of more complicated examples [7]. Cluster analysis is considered the most popular unsupervised ML algorithm. However, research approaches that use the semi-supervised approach have insufficient training data and are

intended only for the normal class, whereas the anomaly class lacks some labels [7]. This generates a minimal number of false alarms and a high detection ratio. Consequently, it is more practical than the supervised mode. However, it is quite challenging to include every anomalous behavior in the training set.

There are two recent important types of ML: reinforcement learning [8] and deep learning [9]. Reinforcement learning algorithms are strongly motivated by behaviorism. They are agents that seek to maximize some reward in a given environment. Deep learning algorithms are a wider class of algorithms that relate directly to artificial neural networks (ANNs). They train machines how to learn by doing what comes naturally to people, or via observation.

This study focuses on the following points:

a) Implementing a DDoS detection platform based on ML to classify the incoming network packet as malicious or benign at the IXP infrastructure.
b) Proposing a research approach to select the essential features using chi-square-based feature extraction.
c) Training and evaluating several effective ML methods based on the selected features using a public DDoS dataset.
d) Selecting the best ML-trained model based on the detection time and classification results for deploying in the SAIXP.

2 Background

Data centers of various sizes and types, as well as the fiber-optic links that link them, comprise the majority of the Internet infrastructure. ISPs are critical for the Internet services of any country [10]. They could become a potential point for an attacker to severely affect the service providers of the Internet. However, they play an important role in mitigating Internet attacks [10,11].

In [12], the IXP is defined as a network facility that enables the interconnection of two or more independent companies of Internet infrastructure, such as ISPs and CDNs. At the core of their infrastructure, an IXP is fundamentally installed in one or more physical places. It has network switches that transfer traffic between the networks of various members. Without IXPs, traffic moving between networks might have to be carried from source to destination via an anonymous intermediary network called a transit provider. The IXP is mainly a layer 2 (open system interconnection (OSI) network model) local area network (LAN) built using one or several ethernet switches that are interconnected together through one or more of the physical buildings. Fundamentally, a home network and an IXP are the same. The difference is only in the scale; IXPs can range from 100 Mbps to many Tbps. Their main objective is to ensure that several network routers are effectively connected. Moreover, a person's house typically has one router and a small number of mobile or computer devices [13].

The denial of service (DoS) and DDoS are the most popular attacks on such networks. Hackers in DoS and DDoS attacks take control of many servers, forming botnets. These botnets are used to send numerous requests using the victim's address as a fake return address. These service requests are invalid and have fake return addresses that mislead the server when it attempts to verify the requestor [14]. When fake requests are continuously handled, the server becomes overloaded. Eventually, the server becomes unavailable for legitimate requests. This is because the attackers organize to send many fake traffic packets to overburden the victim's network connections or computing resources and make the victim nonresponsive/unavailable [15].

The DDoS attacks when launched on an IXP can block the Internet service for the entire or part of that country and will harm the functionality of the victim's organization. These organizations can belong to the government or private sector [16]. DDoS attacks are becoming popular because an increasing number of devices are coming online through the Internet of Things (IoT). IoT devices frequently use common passwords that make them vulnerable to compromise. IoT device infection is frequently ignored by users, and an attacker can easily attack hundreds of thousands of such devices to perform a large-scale attack without the knowledge of device owners.

The cluster of data centers in Saudi Arabia is run and operated mainly by ISPs (namely Saudi Telecoms Company (STC), Mobily, Zain, and Internet Services Unit (ISU)) with other data service providers. Unfortunately, the country has suffered from multiple national cyberattacks that target significant and critical infrastructure (such as Saudi Aramco) known as Shamoon [15,17,18]. This attack kept Saudi Aramco operations in the dark for at least 3 weeks.

The IXP connects the ISP and operates as a hub to link the ISPs with CDNs, such as Google and Facebook. Therefore, it can be considered the primary point of incorporating security policies. Based on this study, if Saudi Arabia implements SAIXP, it would add value from a security viewpoint. It will act as an additional security layer that can be used to mitigate and address most Internet threats at the national level.

3 Literature Review

In general, blackholing is employed as a DDoS mitigation technique both within and between autonomous systems (ASs) [19]. Consequently, the victim's ASs use border gateway protocol (BGP) to notify the upstream network of the attacked target IP prefix. Typically, the traffic that moves toward these prefixes (routing subnets) is dropped upstream at the AS ingress point. This decreases the volume of traffic for every upstream ASs in addition to the destination network. Traditionally, blackholing has been implemented and applied at the edges of AS routers. However, it has gradually moved from the edge (client or network provider) to the Internet's core (IXPs and ISPs) [19].

IXP members usually utilize server routers to distribute BGP announcements to their neighbors [3]. Fig. 1 shows an example of a selective advertisement for specific peers or to all/none. Additionally, IXP members can utilize blackholing techniques to discard any traffic pointed to the victim's prefix (the victim subnet). Hence, the owner advertises its prefix to the server routers with the BGP of the IXP blackholing community. All communities of ASs are known to all IXP members because public information is listed on the portals of the IXPs. The number of ASs that use blackholing universally has increased in the last 3 years, approximately 60% of which depend on the IXP-based variant [3].

Nowadays, ML plays a vital role in detecting DoS and DDoS attack [20] patterns in incoming traffic. Many studies have applied ML algorithms to detect DDoS and other attacks. State-of-the-art findings from the literature are listed below.

Transmission control protocol synchronize (TCP-SYN) and Internet control message protocol (ICMP) flood attacks can be detected using ML methods. KNN and extreme gradient boosting (XGBoost) were used to identify and minimize attack traffic by tracking IP attack sources, whereas normal traffic was almost unaffected [11]. Tuan et al. suggested a DDoS attack reduction in a software-defined network (SDN)-based ISP networks. They implemented the proposed algorithms by deploying a testbed cooperative association for the Internet data analysis (CAIDA) 2007 dataset. The testbed comprises the SDN controller Python OpenFlow controller (POX), which is a networking software platform written in Python, the SDN-enabled switch OpenV Switch (OvS), the usual user, the traffic

replay system, and the victim. It determined the accuracy and discussed the trade-off between accuracy and mitigation performance. The experiments showed good results with over 98% mitigating the attack [21].

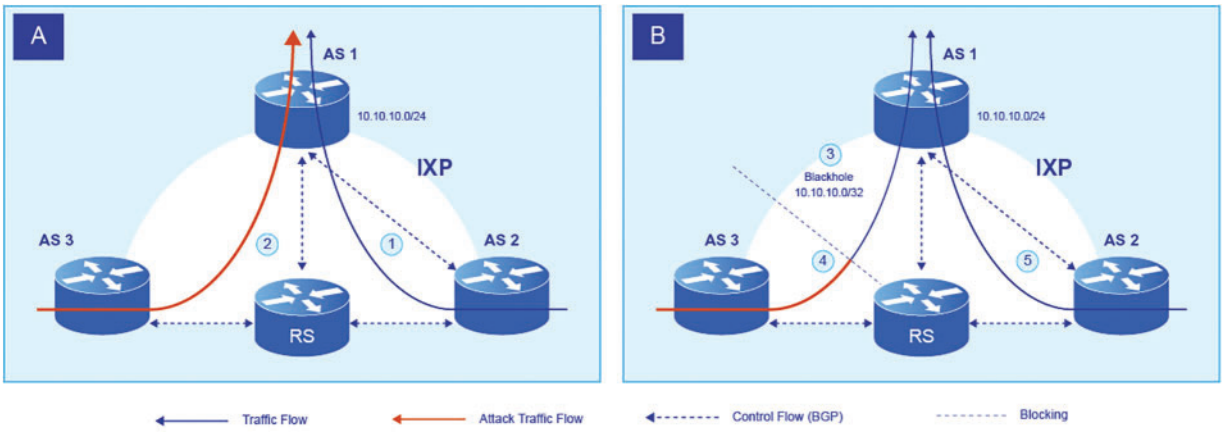


Figure 1: IXP handling the distributed denial of service (DDoS) attack mitigation using the blackholing technique [3]

Lima Filho et al. [22] proposed a DoS detection framework based on ML, which uses the random forest algorithm to identify network traffic directly from network devices based on samples taken from flow protocol traffic. The proposed method achieved results based on previously collected signatures from network traffic samples. The experiments were conducted using three network datasets of DoS raw traffic. The DoS dataset concentrates on the DoS attacks of the application layer mixed with attack-free traces of the ISCXIDS2012 dataset [23]. Eight separate DoS attack assaults from the application layer were produced by four types of attacks using various tools based on CICIDS2017 (the CICIDS2017 dataset was created by ISCX) [24]. The CSE-CIC-IDS2018 dataset is a collective effort between the Communications Security Establishment (CSE) and the Canadian Institute for Cybersecurity (CIC). It includes benign traffic and the most recent common attacks; the new IDS dataset contains seven common updated families of attacks that satisfy real-world criteria and are publicly available [25]. Seven scenarios of attacks are included in the final dataset: heartbleed, brute force, botnet, DoS, web attacks, DDoS, and internal network infiltration. The victim organization comprises 5 departments, 420 machines, and 30 servers in addition to the 50 machines that comprise the attacking infrastructure. Their findings showed that the detection of online attacks was above 96% [22].

Parsaei et al. used four neural network (NN) models to categorize traffic by function. Their approach focused primarily on reducing overhead processing by controllers and classifying network traffic. In the four cases, the proposed approach is evaluated. These are feedforward, multilayer perceptron (MLP), NARX (Levenberg-Marquardt), and NARX (NB). The results were good with more than 95% identification [26].

Bindra et al. used ML algorithms to detect DDoS attacks. They used the knowledge discovery and data mining cup (KDD-CUP) dataset for a detailed investigation of the DDoS attack. The KNN, ID3, NB, and C4.5 algorithms were used and compared. The results were evaluated using the error rate, accuracy, and computation time of the classification algorithms [2].

Pei et al. proposed an ML DDoS attack detection method that involves two steps: extraction of features and model detection. DDoS attack traffic characteristics with a significant proportion were extracted in the feature extraction phase by analyzing the data packets classified based on rules. The extracted features were used as input features of ML during the model detection step. The random forest algorithm was used to train the model for attack detection. The dataset tribe flood network 2000 (TFN2K) was used. Their experimental results showed a good detection rate for the DDoS attack [27].

Polat et al. detected DDoS attacks in the SDN using ML-based models. First, unique features under normal conditions and DDoS attack traffic were obtained from the SDN for the dataset. They created a new dataset using feature selection techniques applied to the existing dataset. The dataset was generated by conducting a simulation for 15 min for each of the sent TCP, user datagram protocol (UDP), and ICMP packets. The dataset contained 65,000 samples of network traffic flow during the DDoS attack. ANN, SVM, NB, and KNN were used as classification algorithms. They showed better results using ML and selection algorithms for detecting DDoS attacks in SDN [28].

Li et al. [29] built a principal component analysis (PCA)-recurrent NN system for detecting DDoS attacks, and they selected most features of the network to identify traffic. It uses the PCA algorithm to reduce characteristic dimensions and also uses dataset KDD (KDD'99 data set created by the defense advanced research projects agency (DARPA) in 1999), which is 9 weeks of network traffic. First, the feature selection was fed into a recurrent NN to train and obtain the classification results. Their model achieved very good classification performance in terms of accuracy, sensitivity, precision, and F-Score [29].

To identify and stop intrusions, Fadel et al. [30] proposed a hybrid deep learning intrusion detection and prevention (HDLIDP) framework that combines signature-based and deep learning NNs. This framework addresses all the aforementioned issues and enhances detection accuracy. Experiments are conducted on datasets from traditional and SDN networks to validate the framework; the results show a significant increase in classification accuracy.

Fadel et al. [31] demonstrated that DDoS attack detection methods can be created and tested on various datasets using the modified whale optimization algorithm (MWOA) feature extraction and hybrid long–short-term memory (LSTM). To decrease prediction errors in the hybrid LSTM algorithm, the weights of the LSTM NN are optimized using the MWOA technique. Using the MWOA-LSTM model, MWOA can also efficiently extract IP packet features and recognize DDoS attacks. Based on precision, recall, and accuracy measurements, the proposed MWOA-LSTM framework outperforms conventional SVM and genetic algorithms (GA), as well as conventional techniques for identifying attacks. Recently, Rusyaidi et al. [32] presented a literature review of the use of ML methods to identify DDoS attacks. In their study, many pertinent research papers were chosen, and they were then evaluated to provide the best performance and supporting data for the applications of ML-based techniques.

For the graph-based solutions to detect the DDoS attacks and communication networks problems, Jiang [33] reviewed a number of studies proposed different models using a graph-based deep learning technique for various problems in communication networks, including both wireless and wired scenarios. The author presented a list of the problems and their solutions for each study and identified the future directions of the research. Li et al. [34] proposed an approach for detecting Distributed Denial of Service (DDoS) attacks using a graph neural network (GNN) model. The authors converted the network traffic into endpoint traffic graphs that contains the information of the relationships of packets and flows. Then, they sent the converted the endpoint traffic graphs to the GNN model to learn accurately the patterns of DDoS attacks. The experimental results showed that the approach

outperforms the state-of-the-art deep learning approaches. Cao et al. [35] proposed a method for detecting DDoS attacks and mitigating their impacts in software-defined networking (SDN) using a spatial-temporal graph convolutional network (ST-GCN) that converts the network into a graph. The switches' state is sensed by the proposed method through sampling with in-band network telemetry (INT). Then, network state is inputted into the ST-GCN model. Finally, the method finds out the DDoS attack flows that pass through the switches. Moreover, the authors mitigated and minimized the impact of DDoS attacks for the traffic on the legitimate network. The results showed that the method detects accurately the path of DDoS attacks flows and improves the accuracy of detection by nearly 10%.

The following are the limitations of the aforementioned studies:

- With the growing demand for IoT and cloud computing, there is a need for more research to mitigate DDoS and investigate attacks that may come from inside, IoT, or the Clouds.
- There is a need to perform more analysis of DDoS attacks based on system vulnerabilities, enhancement of multi-class classification, system self-configuration, development of correlation methods for triggered alarms, formulation of security measures, and application and comparison of more classification algorithms.
- There is a need to investigate DDoS on different device platforms (iOS, Windows, and Linux) and to recognize flows that are of a different application.
- There is a need to apply feature selection to packet datasets and implement the DDoS detection platform in a real network environment [36].

4 Proposed Approach

The proposed approach aims to develop an effective and efficient ML-based DDoS detection model for protecting the SAIXP platform. The approach uses different ML methods, such as DT, SVM, NB, and KNN, to select the most appropriate one from them. Explanations and more details about these ML methods can be read in [37]. Each is evaluated based on two types of justification techniques and uses a representative set of network traffic features containing normal traffic and DDoS attack traffic. The first technique is a holdout technique in which the evaluation dataset is divided into two sets: training and testing sets. Part of the training set is taken for validation in this technique. The second technique is a 10-fold cross validation technique. In this technique, the dataset is divided into 10 parts: nine parts for training and one part for testing. This partitioning process is repeated 10 times. In both techniques, the proposed approach contributes to selecting few features using a chi-square method without decreasing the accuracy results for efficiency and effectiveness. The research approach proposes deploying the trained ML in the IXP, as shown in Fig. 2. Internet infrastructure providers, such as ISPs and CDNs, physically connect with one another at the IXP. These areas, which are on the "edge" of many networks, enable network providers to share transit outside of their network. Companies that have a presence inside an IXP location can minimize their path to transit from other participating networks, reducing latency, improving round-trip time, and possibly lowering expenses for themselves, other ISPs and their clients, and ISPs transport Internet traffic. Some ISPs have a wide geographic reach and are huge. Some are smaller and only exist in a single nation or continent. The result is a 3-tier ISP model. These ISPs serve as the Internet's skeleton

at tier 1. By expanding their network internationally, they have a global presence. Internet transit is offered to customers by tier 1 ISPs through private peering agreements with other tier 1 ISPs. A tier 1 ISP frequently assigns multiple autonomous system networks (ASNs), each of which is used for a distinct function and service. These ISPs utilize peering with other tier 2 ISPs and pay Internet transit from tier 1 ISPs to form tier 2. Typically, they are national providers. In tier 3, these ISPs only purchase Internet transit from tier 2 or tier 3 ISPs to provide Internet connectivity to their end customers. They primarily give their clients local access to the Internet. They may be local or regional (such as cities or metros) providers. The flowcharts of the two techniques used in the research approach are given in Figs. 3 and 4.

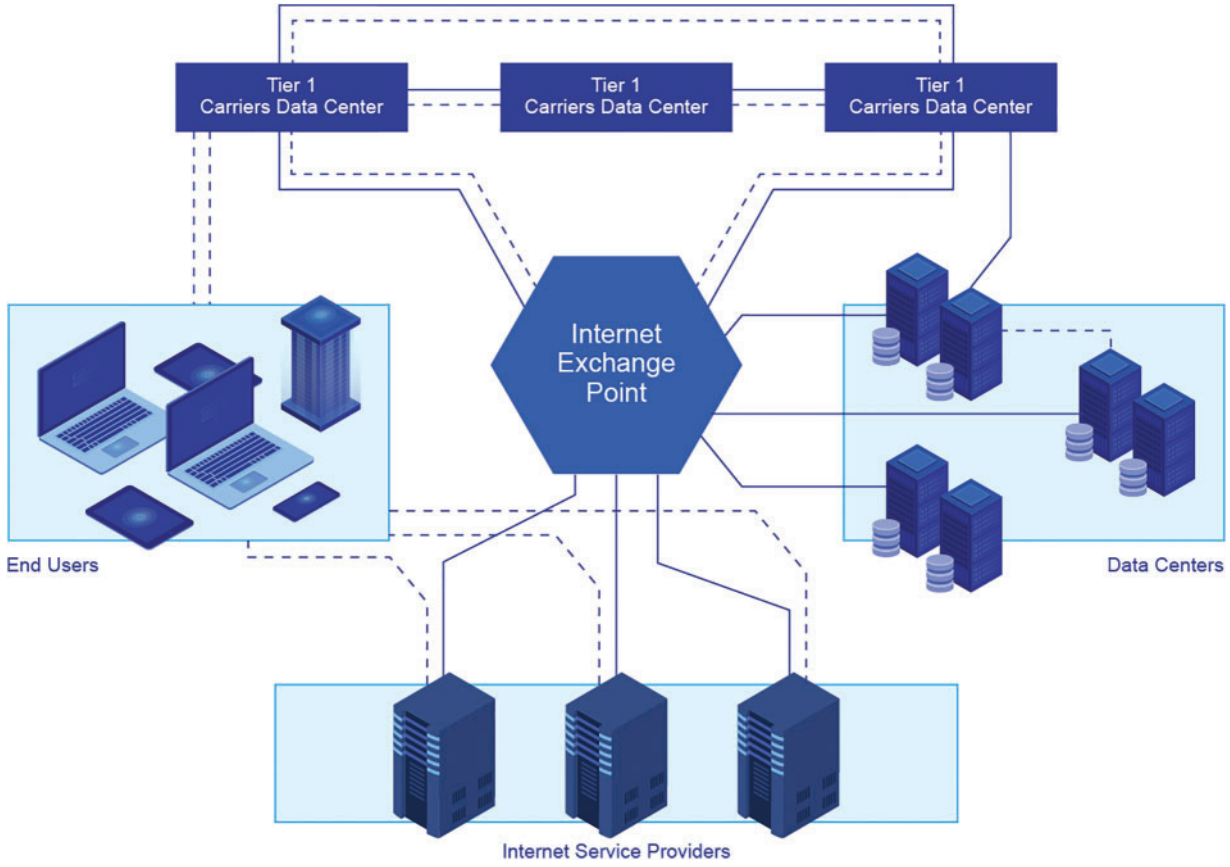


Figure 2: Deployment place of the proposed research approach

Each flowchart contains a set of steps to extract, clean, normalize data network features, select, train, and evaluate ML methods. Finally, the best method is recommended for deployment on the developed platform.

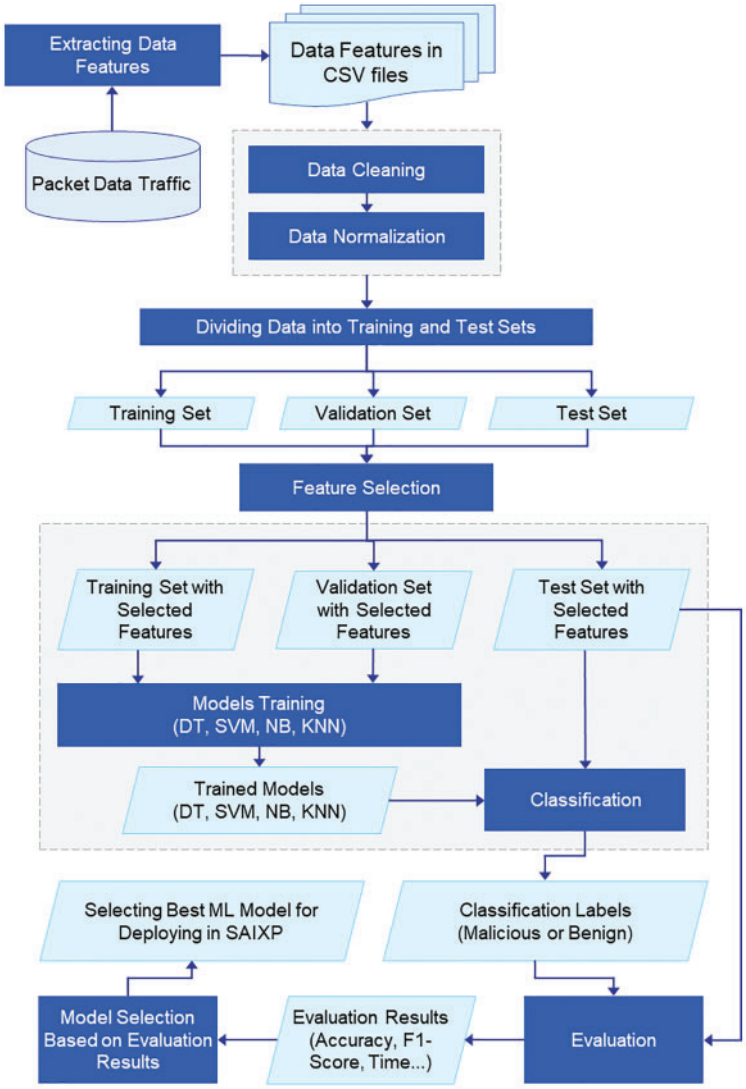


Figure 3: The flowchart of the first technique of the research approach

The research approach starts by extracting the data of important features from the packet data traffic in the data store and saving them in comma-separated value (CSV) files. Then, the data cleaning and feature selection steps are executed. The data cleaning step is used to remove nonfinite and null values to obtain data ready for training the ML models. Additionally, in this step, duplicated columns are processed by dropping the second occurrence. The labels of the data traffic are encoded into numbers to enable the classifier to learn the class number to which each tuple belongs, and this is called label encoding.

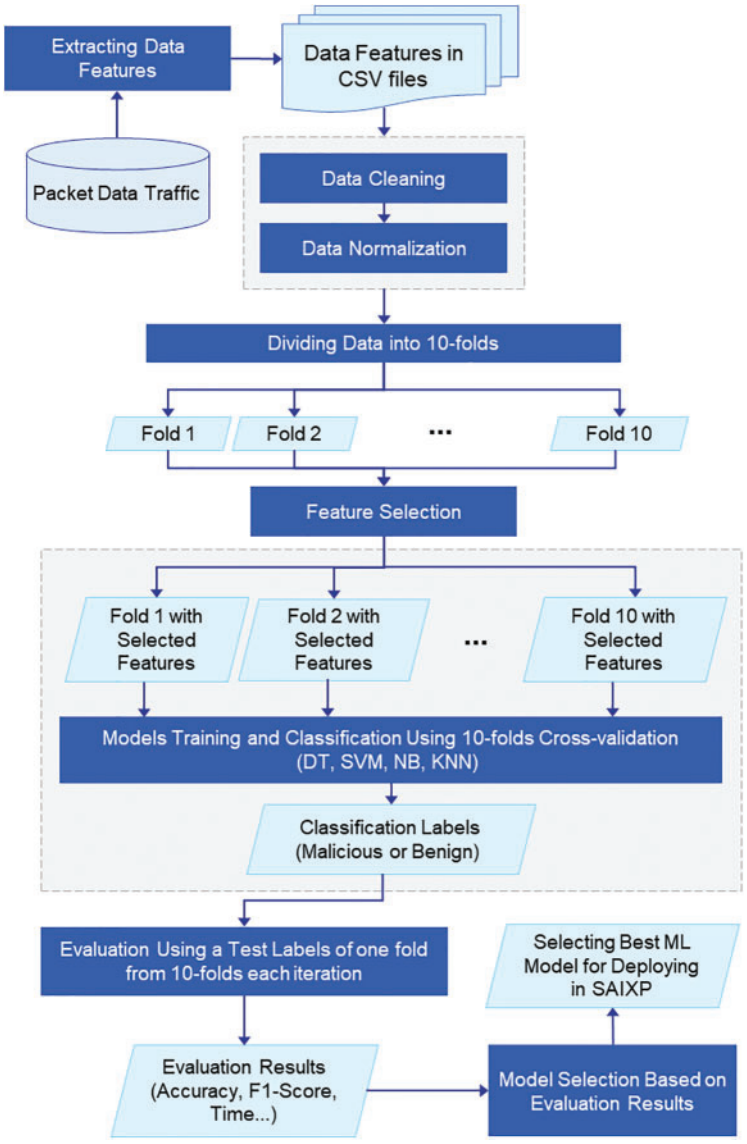


Figure 4: The flowchart of the second technique of the research approach

The numerical data in the dataset have various ranges, causing the classifier some challenges during training to compensate for these differences. Consequently, it is important to normalize these values in each feature where the maximum value is one and the minimum value is zero. This gives more homogeneous values to the classifier and maintains relativity between the values of each attribute. Data normalization using the min–max technique is applied. Subsequently, the features in the dataset, which have high chi-squared statistics, are selected to increase the score and performance of the proposed ML model. A chi-square-based technique is used for feature selection because it is easier to compute than some other techniques and does not need any assumption regarding the distribution of feature values than statistics techniques that assume a certain characteristic about the distribution of the feature values [38]. To select the important features with the best chi-square metric, two parameters are passed to the method: one is the scoring metric, which is chi-squared, and the other is the value of K, which

signifies the desired number of features in the final dataset. Then, using only the selected features, the training and evaluation processes of the ML models are performed, and the best model is deployed in the SAIXP. Algorithm 1 describes the main steps of the proposed approach.

Algorithm 1: Main steps of the proposed research approach

1. Extracting data features.
2. Cleaning data features.
3. Normalizing data features using a min–max technique.
4. Dividing dataset into training, validation, and testing sets using holdout and 10-fold cross-validation techniques.
5. Selecting the most important features using a chi-square-based feature selection technique.
6. Training various ML models.
7. Classifying test set instances using trained ML models.
8. Evaluating classification results of test set instances using evaluation metrics.
9. Selecting the best ML model based on the evaluation results.

5 Experiments and Discussion

This section validates the proposed approach for the early warning and protection of the SAIXP platform. The ML-based approach can detect DDoS attacks from benign traffic packets using an effective and efficient model. Through experiments, the feature selection and ML methods mentioned in the research approach are applied to develop the proposed platform. The dataset adopted for evaluation along with evaluation metrics and the results is explained in the next subsections. Two experiments are conducted for evaluation and validation. The first experiment is based on the holdout evaluation technique, in which the dataset is divided into three sets: training, validation, and test sets. The second experiment is performed using a 10-fold cross validation technique, in which the dataset is split into 10 sets: one set for testing and the remaining 9 sets for training. This experiment is run 10 times, and at each time, one different set is used for testing. In both experiments, the hyperparameters of the ML models are initialized to have the default values. The experiments are conducted on a laptop Core i7 CPU 2.20 GHz, RAM 32 GB, Display Card 8 GB NVIDIA GeForce GTX 1070, and Windows 10 Operating System.

5.1 Dataset

The Canadian Institute for Cybersecurity 2017 (CICIDS2017) dataset is selected for evaluating the proposed models because it is more recent, includes more features, and contains DDoS attacks as well as a large set of instances for better learning. It was produced by the Canadian Institute for Cybersecurity. A novel systematic approach was proposed by defining two types of profiles to create this valid dataset. The dataset contains various up-to-date multistage attacks, such as Heartbleed, DDoS, and different types of DoS attacks. Moreover, a diversity of modern protocols is involved. The dataset has 80 columns for each Netflow record saved in eight files of CSV formats, thereby making it easy to import for training the ML methods [39]. These files contain normal traffic named benign traffic and malicious traffic, which are different types of attacks. There are 14 types of attacks in this dataset, as presented in Table 1.

The data are contained in the eight CSV files, and each of them contains different attack data at different times. First, all data from all files are merged into one data CSV file with a size of 2,830,743 instances with 79 features and a label field.

Table 1: Summary of the CICIDS2017 dataset

File name	Type of traffic	Number of record
Monday- WorkingHours.pcap_ISCX.csv	Benign	529,918
Tuesday-WorkingHours.pcap_ISCX.cSV	Benign	432,074
	SSH-Patator	5,897
	FTP-Patator	7,938
Wednesday-WorkingHours.pcap_ISCX.cSV	Benign	440,031
	DoS Hulk	231,073
	DoS GoldenEye	10,293
	DoS Slowloris	5,796
	DoS Slowhttptest	5,499
	Heartbleed	11
Thursday-WorkingHours-Morning-WebAttacks.pcap_ISCX.csv	Benign	168,186
	Web Attack- Brute Force	1,507
	Web Attack-SQL Injection	21
	Web Attack-XSS	652
Thursday-WorkingHours-Afternoon-Infilteration.pcap_ISCX.cSv	Benign	288,566
Friday-WorkingHours-Moming.pcap_ISCX.csv	Infiltration	36
	Benign	189,067
	Bot	1,966
Friday-WorkingHours-Afternoon-PortScan.pcap_ISCX.csv	Benign	127,537
	Portscan	158,930
Friday-WorkingHours-Afternoon-DDos.pcap_ISCX.csv	Benign	97,718
	DDoS	128,027
Total Instance/Record		2,830,743

Because the scope of these is to detect DDoS attacks from normal traffic, the instances of DDoS and benign network traffic are selected to generate the adopted DDoS dataset for evaluating the ML methods of the SAIXP platform. The distribution of instances in the DDoS dataset is shown in Fig. 5.

5.2 Evaluation Metrics

The evaluation metrics used to measure the results of the proposed models are accuracy (ACC), precision (PR), recall (RE), f1-score, and false alarm rate (FAR). They are selected because they produce comparable results and are frequently used to evaluate models in the ML field. These evaluation metrics can be defined as follows:

$$\text{Accuracy (ACC)} = (TP + TN)/(TP + FP + TN + FN) \quad (1)$$

$$\text{Precision (PR)} = \text{TP}/(\text{TP} + \text{FP}) \tag{2}$$

$$\text{Recall (RE) or True Positive Rate (TPR)} = \text{TP}/(\text{TP} + \text{FN}) \tag{3}$$

$$\text{F1–Score} = 2 * ((\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall})) \tag{4}$$

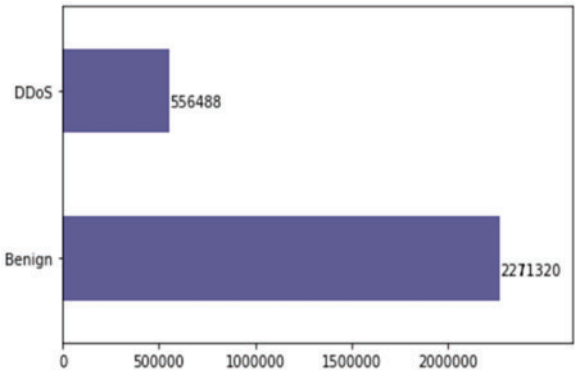
$$\text{False Positive Rate (FPR)} = \text{FP}/(\text{FP} + \text{TN}) \tag{5}$$

$$\text{False Negative Rate (FNR)} = \text{FN}/(\text{FN} + \text{TP}) \tag{6}$$

$$\text{True Negative Rate (TNR)(Specificity)} = \text{TN}/(\text{TN} + \text{FP}) \tag{7}$$

$$\text{False Alarm Rate (FAR)} = (\text{FPR} + \text{FNR})/2 \tag{8}$$

$$\text{Detection Rate (DR)} = \text{No. of Detected Attacks/Total No. of Attacks} \tag{9}$$



Benign	2271320
DDoS	128025

Figure 5: The distribution of instances in the distributed denial of service (DDoS) dataset

where TP denotes a true positive sample correctly classified by the model, TN denotes a true negative sample correctly classified by the model, FP denotes a false positive in which a negative sample is wrongly classified as positive, and FN denotes a false negative in which a positive sample is wrongly classified as negative. To verify the effectiveness of the proposed model for developing the SAIXP platform, it is compared with other methods in previous studies based on the same datasets used.

5.3 Results

After preparing the values of extracted features from the network data traffic in CSV files and processing them, features with high chi-squared statistics are selected in the feature selection step to improve the performance of ML models. The fit transform function is used to transform the feature values of the dataset. Consequently, the final dataset with the desired selected features is taken. The score associated with each feature is plotted, as depicted in Fig. 6. In addition, Fig. 7 illustrates that the importance of the features are sorted based on the percentage of their cumulative scores.

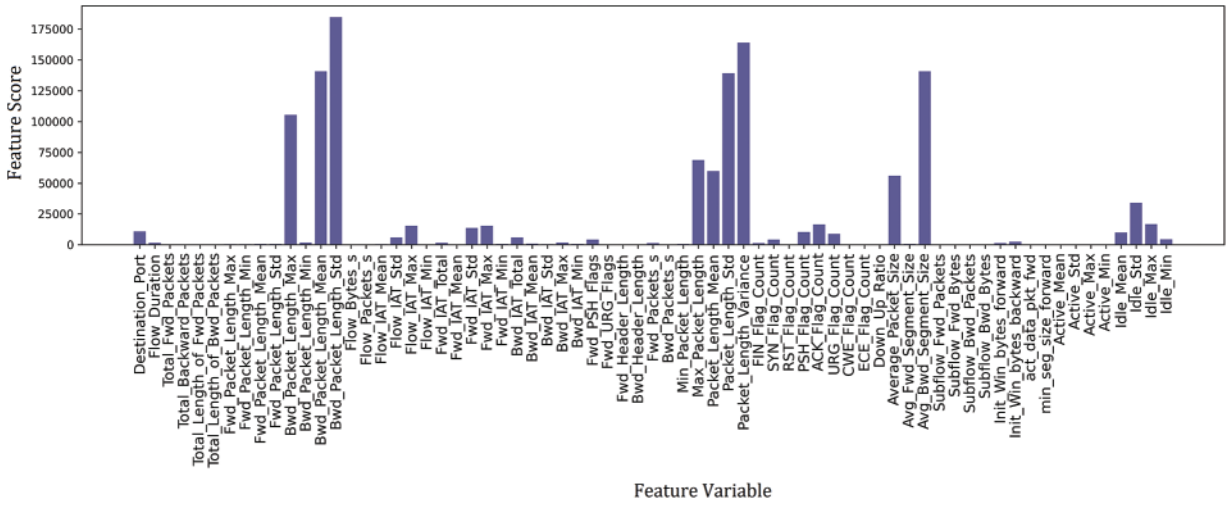


Figure 6: The scores associated with each feature in the distributed denial of service (DDoS) dataset

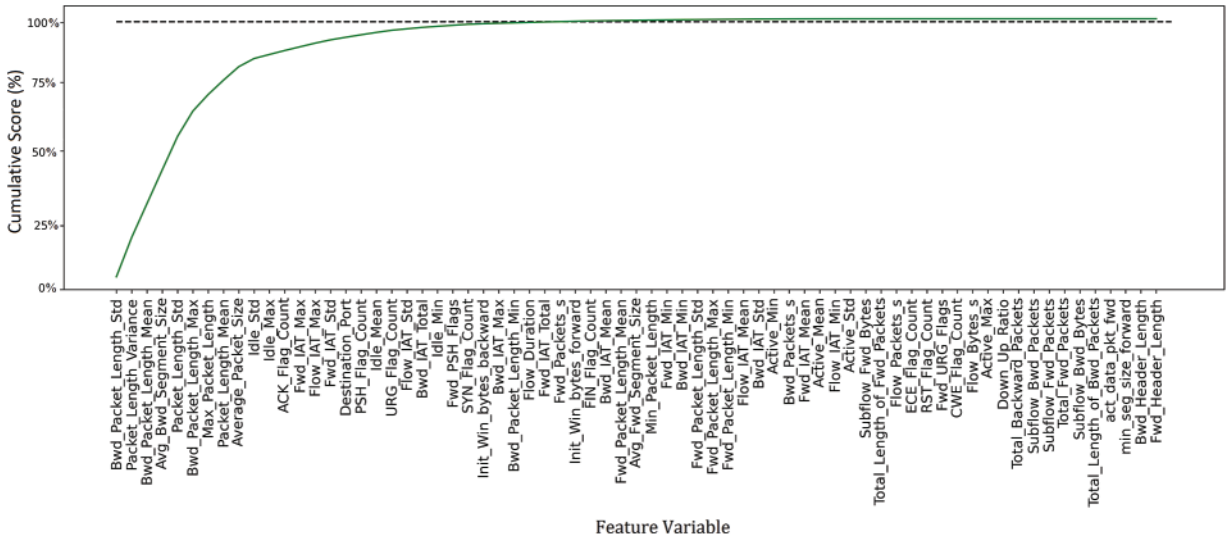


Figure 7: Sorting the features of the distributed denial of service (DDoS) dataset based on the percentage of their cumulative scores

From Fig. 7, it is obvious that 99.99% of the information is contained in the first 20 features. Therefore, this is the cutoff point to perform feature selection with K = 5, 10, 15, and 20 best features and evaluate the performance results. According to the two techniques (holdout and 10-fold cross-validation) utilized in the research approach, the validation and testing results of the ML algorithms are presented in two experiments.

The following subsection explains the results of each experiment in detail. The experimental results are measured using the evaluation metrics presented in Subsection 5.2. To assess the efficiency of the proposed model, training and testing times are computed during the first experiment.

5.3.1 Experiment 1

This experiment performs a holdout evaluation technique. The holdout technique splits the DDoS dataset into three sets with ratios of 60%, 20%, and 20% for training, validating, and testing, respectively. Table 2 presents the number of instances in each set.

Table 2: Number of instances in experiment 1

Class label	Training set	Validation set	Testing set	Total
BENIGN	1362792	454264	454264	2271320
DDoS	76815	25605	25605	128025
Total	1439607	479869	479869	2399345

To select the accurate ML model with a low number of features as a proposed model for developing the desired platform, these experimental results at different numbers of features less than the optimum number of features, which is 20, are computed using the feature selection method explained in the proposed approach section (Section 4). Here, we train the adopted ML models on the training set with the first 5, 10, 15, and 20 features that have the best chi-square scores. Table 3 presents the results of the validation and testing accuracy of the trained models. To show the DT model with the 10 selected features and allow the readers to see the features that identify DDoS attacks, Fig. 8 demonstrates the indices of features, which are the names shown in Fig. 7 and are ordered from left to right.

Table 3: The results of validation and testing accuracy in experiment 1

Classification method	Number of features	Validation accuracy	Testing accuracy
DT	5	0.9806	0.9807
	10	0.9997	0.9998
	15	0.9998	0.9998
	20	0.9999	0.9999
SVM	5	0.9787	0.9787
	10	0.9799	0.9800
	15	0.9790	0.9790
	20	0.9797	0.9797
NB	5	0.9466	0.9466
	10	0.9466	0.9466
	15	0.9566	0.9566
	20	0.9734	0.9736
KNN	5	0.9806	0.9807
	10	0.9858	0.9856
	15	0.9998	0.9998
	20	0.9999	0.9999

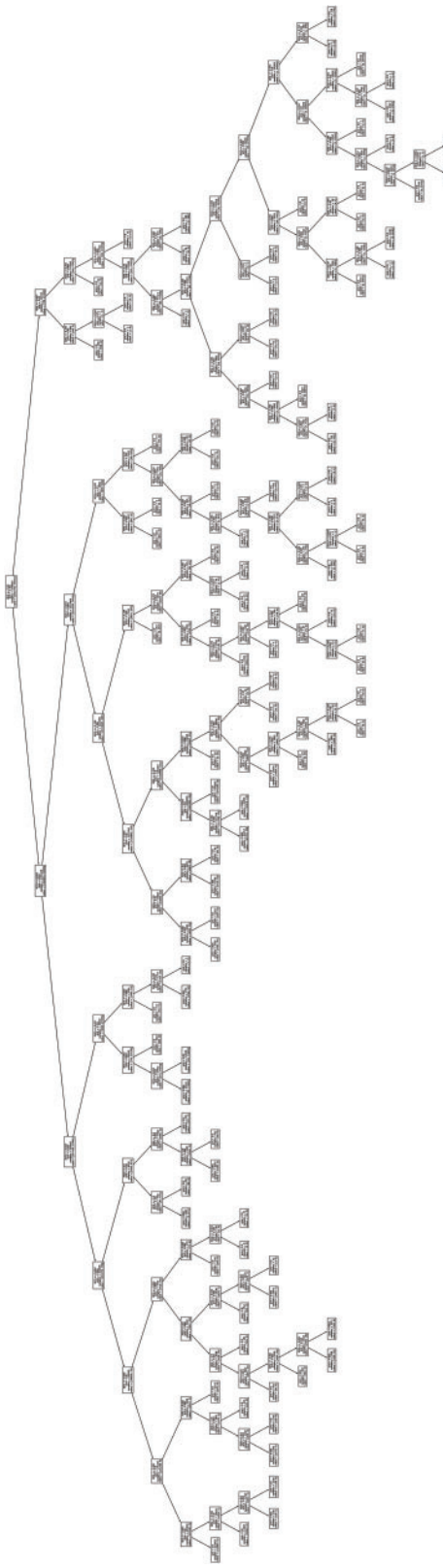


Figure 8: Visualization of the decision tree (DT) model with the indices of the 10 selected features, which are their names depicted in Fig. 7 and are ordered from left to right

Additionally, to show and analyze the performance of ML models to classify DDoS for benign network traffic at all classification thresholds, Table 4 lists the area under curve (AUC) values of the receiver operating characteristic (ROC) for all ML models using the selected features. The AUC is a major evaluation metric for measuring the performance of classification models at various threshold values. It is a probability curve that gives the quantity or degree of separability between classification classes. The AUC states how much the model can differentiate between network traffic classes. Higher AUC values mean that the model can classify benign and DDoS classes as benign and DDoS, respectively. The ROC curve is plotted with the true positive rate (TPR) against the FPR, where the FPR is on the *x-axis* and the TPR is on the *y-axis*, as shown in Figs. 9e–9h.

Table 4: The area under curve (AUC) values of the testing sets for all machine learning models in experiment 1

Number of features	AUC			
	DT	SVM	NB	KNN
5	0.8191	0.8096	0.5	0.8191
10	0.9991	0.8144	0.50	0.908
15	0.999	0.8141	0.5929	0.999
20	0.9995	0.8191	0.7538	0.9997

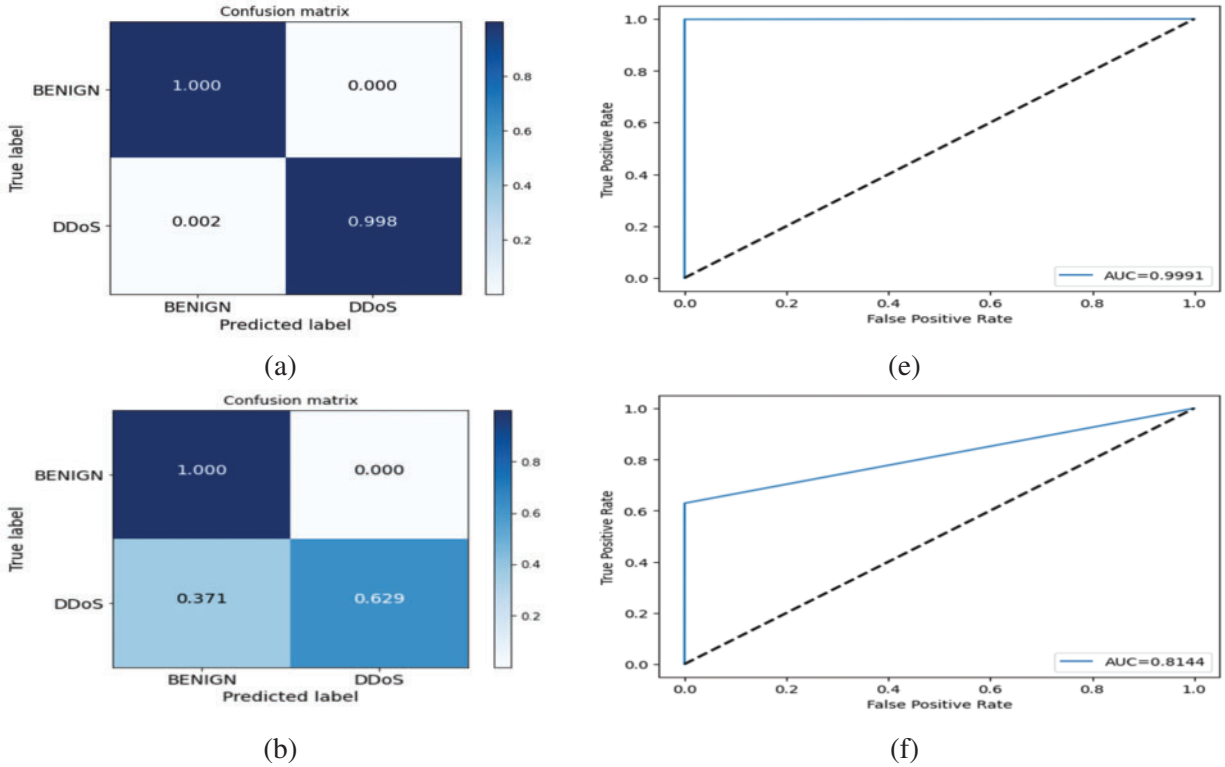


Figure 9: (Continued)

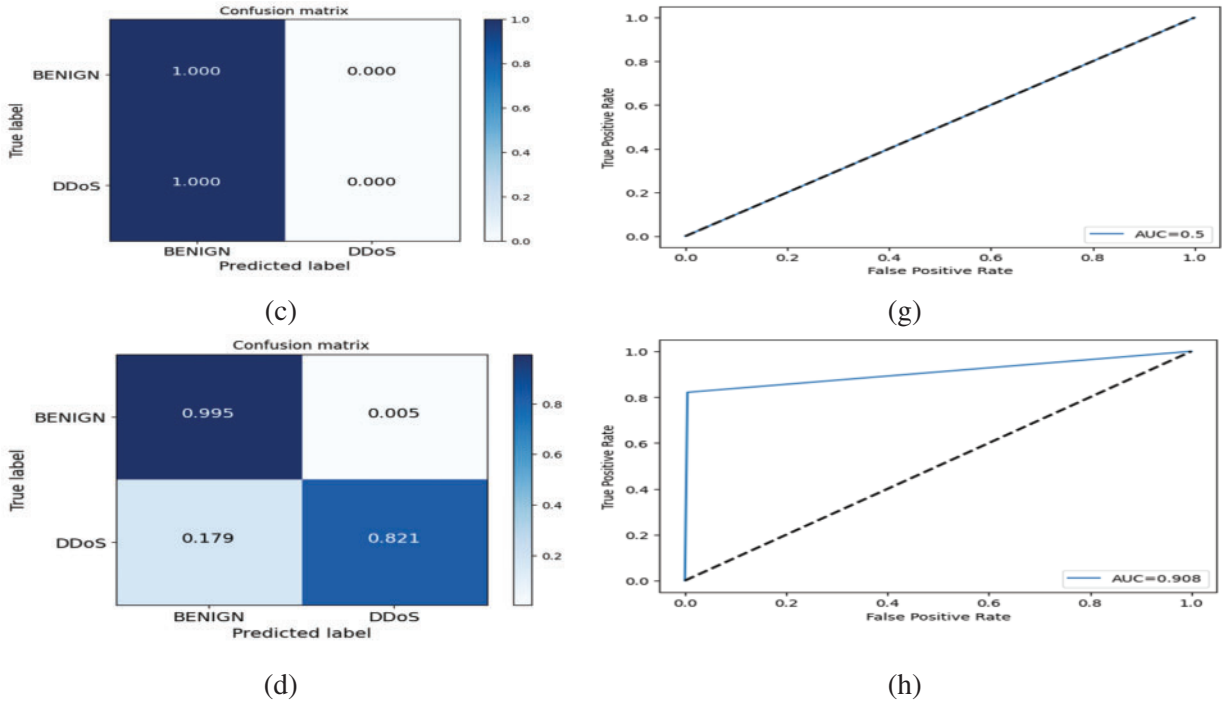


Figure 9: Confusion matrices and plots of AUC for all ML models in experiment 1: (a)–(d) confusion matrices of DT, SVM, NB, and KNN, respectively; (e)–(h) plots of AUC for DT, SVM, NB, and KNN, respectively

As shown in Tables 3 and 4, we can see that the best number of features is 10 and the accurate model is DT. Because computational time is an important factor in efficiency, Table 5 presents the execution time of training and testing in seconds for all the models with all selected features.

Table 5: The execution time of training and testing in seconds of machine learning models in experiment 1

Classification method	Number of features	Training time	Testing time
DT	5	5.748	0.033
	10	6.697	0.033
	15	9.623	0.037
	20	13.484	0.044
SVM	5	16.315	0.010
	10	20.550	0.010
	15	26.566	0.020
	20	28.498	0.024
NB	5	0.022	0.021
	10	0.189	0.028
	15	0.192	0.036

(Continued)

Table 5 (continued)

Classification method	Number of features	Training time	Testing time
	20	0.245	0.281
KNN	5	1511.770	119.284
	10	1627.556	306.444
	15	1988.795	555.745
	20	2208.398	834.686

As shown in Table 5, the execution time of 10 features is more appropriate than other sets of features based on the attained performance results in Tables 3 and 4. In addition, it can be observed that the NB model has the lowest execution time for training and testing, and the DT model has the second lowest execution time. The DT model achieves a higher accuracy result than the NB model. Therefore, the DT model is selected with the first 10 features obtained from the chi-square feature selection method as a proposed ML model to develop the SAIXP platform.

To validate this selection, more analysis of the experimental results is given by comparing the other evaluation metrics and through the 10-fold cross validation technique in the second experiment between all ML models and using the 10 selected features. Tables 6–10 shows the results of evaluation metrics, and Fig. 9 demonstrates the confusion matrices and plots of AUC for all ML models in experiment 1.

Table 6: The test results of precision, recall, and F1-score of the decision tree (DT) in experiment 1

Class label	Precision	Recall	F1-score	DR
BENIGN	0.9999	0.9998	0.9999	0.9981
DDoS	0.9972	0.9983	0.9977	
Macro avg.	0.9985	0.9991	0.9988	
Weighted avg.	0.9998	0.9998	0.9998	

Table 7: The test results of precision, recall, and F1-score of SVM in experiment 1

Class label	Precision	Recall	F1-score	DR
BENIGN	0.9795	0.9998	0.9895	0.6289
DDoS	0.9936	0.6289	0.7703	
Macro avg.	0.9865	0.8144	0.8799	
Weighted avg.	0.9803	0.9800	0.9778	

5.3.2 Experiment 2

The second experiment is used to validate the results of the first experiment through the 10-fold cross validation technique. This validation technique is a comprehensive evaluation method because it covers all dataset examples in the testing process. Table 11 demonstrates the average results of

precision, recall, F1-score, and accuracy metrics for the 10-fold test sets, and Table 12 presents the average results of TNR, FPR, FNR, and FAR metrics for the 10-fold cross validation runs.

Table 8: The test results of precision, recall, and F1-score of NB in experiment 1

Class label	Precision	Recall	F1-score	DR
BENIGN	0.9466	1.0000	0.9726	0
DDoS	0.0000	0.0000	0.0000	
Macro avg.	0.4733	0.5000	0.4863	
Weighted avg.	0.8961	0.9466	0.9207	

Table 9: The test results of precision, recall, and F1-score of KNN in experiment 1

Class label	Precision	Recall	F1-score	DR
BENIGN	0.9900	0.9949	0.9924	0.8212
DDoS	0.9001	0.8212	0.8589	
Macro avg.	0.9451	0.9080	0.9256	
Weighted avg.	0.9852	0.9856	0.9853	

Table 10: The test results of TNR, FPR, FNR, and FAR of all models in experiment 1

Classification model	Metric	BENIGN	DDoS
DT	TNR	0.998	1.000
	FPR	0.002	0.000
	FNR	0.000	0.002
	FAR	0.001	0.001
SVM	TNR	0.629	1.000
	FPR	0.3711	0.0002
	FNR	0.0002	0.3711
	FAR	0.186	0.186
NB	TNR	0.000	1.000
	FPR	1.000	0.000
	FNR	0.000	1.000
	FAR	0.5	0.5
KNN	TNR	0.821	0.995
	FPR	0.179	0.005
	FNR	0.005	0.179
	FAR	0.092	0.092

Table 11: The average results of precision, recall, F1-score, and accuracy for the 10-fold test sets in experiment 2

Class label	Evaluation metric	Classification method			
		DT	SVM	NB	KNN
BENIGN	Precision	0.9999	0.9796	0.9466	0.9999
	Recall	0.9998	0.9998	1.0000	0.9998
	F1-score	0.9999	0.9896	0.9726	0.9999
DDoS	Precision	0.9972	0.9942	0.0000	0.9971
	Recall	0.9983	0.6302	0.0000	0.9982
	F1-score	0.9977	0.7714	0.0000	0.9977
Both	Accuracy	0.9998	0.9801	0.9466	0.9998

Table 12: The average results of TNR, FPR, FNR, and FAR for the 10-fold test sets in experiment 2

Class label	Evaluation metric	Classification method			
		DT	SVM	NB	KNN
BENIGN	TNR	0.9983	0.6301	0.0000	0.9982
	FPR	0.0017	0.3699	1.0000	0.0018
	FNR	0.0002	0.0002	0.0000	0.0002
	FAR	0.00095	0.18505	0.5	0.001
DDoS	TNR	0.9999	0.9998	1.0000	0.9998
	FPR	0.0002	0.0002	0.0000	0.0002
	FNR	0.0017	0.3699	1.0000	0.0018
	FAR	0.00095	0.18505	0.5	0.001

As depicted in Tables 11 and 12, it can be observed that the results of DT are still higher than those of the other ML models, particularly in terms of precision, recall, FNR, and FAR for detecting DDoS attacks. The results of the KNN model seem to be competitive with the results of DT. However, the DT model is more efficient in terms of the computational time for training and detection, which represents the efficiency factor. It can be observed that the recall value of NB for the benign class is one, which means that the NB model can detect all instances of that class and does not detect DDoS instances. It detects DDoS attacks as benign instances. The NB model suffers from a higher bias problem than the other models.

5.4 Comparison Results

To benchmark the performance of the proposed ML model for detecting DDoS attacks, the results of DR, FAR, and precision are compared with a recent-related study proposed by Limo Filho et al. [22], as shown in Table 13.

Table 13: The average results of TNR, FPR, FNR, and FAR for the 10-fold test sets in experiment 2

Works	Dataset	Number of features	DR	FAR	Precision
Limo Filho et al. [22]	CIC-DoS	20	0.9360	0.0004	0.9990
	CICIDS2017		0.8000	0.0020	0.9920
This study	CICIDS2017-DDoS	10	0.9981	0.00095	0.9998

As depicted in Table 13, the proposed model achieves a higher DR than that of the related study. Moreover, the competitive results of FAR and precision are attained by the proposed approach with a small number of features. Therefore, the DT model with the proposed selected features should be more effective and efficient for detecting DDoS attacks from normal network traffic.

6 Conclusions and Future Work

Because all Internet traffic entering a country should pass through the IXP, it is vital to perform malicious traffic analysis at that point. In this study, an effective and efficient ML-based DDoS early detection approach is proposed to protect the SAIXP platform. The DDoS attacks are becoming a more serious threat daily. They created massive numbers of infected machines known as botnets. The proposed approach contains some steps starting with extracting data features in CSV files from packet data traffic in the data store and then cleaning the extracted data, performing data normalization, dividing data into sets using holdout and 10-folds, selecting important features, training the ML models, classifying data features using trained models, evaluating the results, and selecting the best model for deployment in the SAIXP platform. The effectiveness and efficiency of the proposed approach result from selecting an accurate ML model with few input features. The ML models are tested using holdout and 10-fold tests on a public large-size dataset. The experiments showed that the performance of the DT classifier achieved a high accuracy result (99.98%) with few features (10 features). The experimental results confirm the applicability of using DT and chi-square for DDoS detection and early warning in SAIXP.

However, it is crucial to acknowledge potential limitations and areas for future research:

1. **Adapting to evolving DDoS attack strategies:** As attackers continue to develop new techniques, the proposed model must be adaptable and capable of detecting emerging DDoS attack patterns.
2. **Real-time performance optimization:** Ensuring that the proposed approach can detect and mitigate DDoS attacks in real-time is essential for the effective protection of the SAIXP platform.
3. **Model robustness against adversarial attacks:** Investigating the resilience of the DT classifier against adversarial attacks and evasion techniques will be crucial for maintaining reliable DDoS detection.
4. **Validation with additional datasets and IXPs:** Further validation of the proposed approach on different datasets and in the context of other IXPs will help to assess its generalizability and applicability in broader settings.

In conclusion, the proposed ML-based DDoS early detection approach shows promise in addressing the critical challenge of protecting the SAIXP platform from DDoS attacks. By building on the findings of this study and addressing the identified limitations, we can continue to advance DDoS mitigation efforts and contribute to a more secure and resilient internet infrastructure.

Acknowledgement: We would like to thank the reviewers for their time and effort in reviewing this manuscript. We honestly appreciate all the valuable suggestions and comments that have assisted us in improving the quality of this manuscript.

Funding Statement: The authors received no specific funding for this study.

Availability of Data and Materials: The data supported the findings of this study are available upon request from the corresponding author.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] S. Rao, A. K. Verma and T. Bhatia, "Evolving cyber threats, combating techniques, and open issues in online social networks," in *Handbook of Research on Cyber Crime and Information Privacy*, Hershey, PA, USA: IGI Global, pp. 219–235, 2021.

[2] N. Bindra and M. Sood, "Detecting ddos attacks using machine learning techniques and contemporary intrusion detection dataset," *Automatic Control Computer Sciences*, vol. 53, no. 5, pp. 419–428, 2019.

[3] C. Dietzel, M. Wichtlhuber, G. Smaragdakis and A. Feldmann, "Stellar: Network attack mitigation using advanced blackholing," in *Proc. of the 14th Int. Conf. on Emerging Networking Experiments and Technologies*, Heraklion, Greece, pp. 152–164, 2018.

[4] N. Chatzis, G. Smaragdakis, J. Böttger, T. Krenc and A. Feldmann, "On the benefits of using a large ixp as an internet vantage point," in *Proc. of the 2013 Conf. on Internet Measurement Conf.*, Barcelona, Spain, pp. 333–346, 2013.

[5] D. Kobialka, "Kaspersky lab study: Average cost of enterprise ddos attack totals $2M," *MSSP Alert*, 2018. https://www.msspalert.com/cybersecurity-research/kasperskylab-study-average-cost-of-enterprise-ddos-attack-totals-2m/

[6] C. Buragohain, M. J. Kalita, S. Singh and D. K. Bhattacharyya, "Anomaly based ddos attack detection," *International Journal of Computer Applications*, vol. 123, no. 17, pp. 35–40, 2015.

[7] G. Nadiammai and M. Hemalatha, "Effective approach toward intrusion detection system using data mining techniques," *Egyptian Informatics Journal*, vol. 15, no. 1, pp. 37–50, 2014.

[8] K. Rungta, *Tensorflow in 1 day: Make your own neural network*. Packt Publishing, 2019.

[9] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks*, vol. 61, no. 3, pp. 85–117, 2015.

[10] B. Rowe, D. Reeves and M. Gallaher, "The role of internet service providers in cyber security," *Computers & Security*, vol. 28, no. 1, pp. 1–11, 2009. https://doi.org/10.1016/j.cose.2008.09.002

[11] M. Van Eeten, J. M. Bauer, H. Asghari, S. Tabatabaie and D. Rand, "The role of internet service providers in botnet mitigation an empirical analysis based on spam data," *Telecommunications Policy Research Conference*, pp. 1–21, 2010. https://doi.org/10.1016/j.telpol.2010.07.003

[12] M. Thompson, "Understanding physical internet infrastructure vulnerabilities," *The CIP Report*, 2016.

[13] J. Boulmetis and P. Dutwin, *The ABCs of Evaluation: Timeless Techniques for Program and Project Managers*. New York, US: John Wiley & Sons, 2014.

[14] B. Liu, *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Berlin, Germany: Springer, 2011.

[15] D. S. Bujud, A. S. Al Ghamdi and M. N. Saqib, "A survey of layered approach to threats and countermeasures," in *Proc. of the Int. Conf. on Security and Management (SAM)*, Athens, Greece, pp. 189–194, 2019.
[16] J. M. Estevez-Tapiador, P. Garcia-Teodoro and J. E. Diaz-Verdejo, "Anomaly detection methods in wired networks: A survey and taxonomy," *Computer Communications*, vol. 27, no. 16, pp. 1569–1584, 2004.
[17] S. Alelyani and H. Kumar, "Overview of cyberattack on Saudi organizations," *Journal of Information Security and Cybercrimes Research*, vol. 1, no. 1, pp. 42–50, 2018.
[18] R. A. Al-Mulhim, L. A. Al-Zamil and F. M. Al-Dossary, "Cyber-attacks on Saudi Arabia environment," *International Journal of Computer Networks Communications Security*, vol. 8, no. 3, pp. 26–31, 2020.
[19] C. Dietzel, A. Feldmann and T. King, "Blackholing at ixps: On the effectiveness of ddos mitigation in the wild," in *Int. Conf. on Passive and Active Network Measurement*, Cham, Switzerland, pp. 319–332, 2016.
[20] M. Alenezi and M. J. Reed, "Methodologies for detecting dos/ddos attacks against network servers," in *the Seventh Int. Conf. on Systems and Networks Communications ICSNC*, Lisbon, Portugal, pp. 92–98, 2012.
[21] N. N. Tuan, P. H. Hung, N. D. Nghia, N. V. Tho, T. V. Phan *et al.,* "A ddos attack mitigation scheme in isp networks using machine learning based on SDN," *Electronics*, vol. 9, no. 3, 413, 2020.
[22] F. S. d. Lima Filho, F. A. Silveira, A. de Medeiros Brito Junior, G. Vargas-Solar and L. F. Silveira, "Smart detection: An online approach for dos/ddos attack detection using machine learning," *Security Communication Networks*, vol. 2019, 1574749, 2019.
[23] A. Shiravi, H. Shiravi, M. Tavallaee and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," *Computers Security*, vol. 31, no. 3, pp. 357–374, 2012.
[24] I. Sharafaldin, A. Habibi Lashkari and A. A. Ghorbani, "A detailed analysis of the cicids2017 data set," in *Int. Conf. on Information Systems Security and Privacy*, Cham, Switzerland, pp. 172–188, 2019.
[25] J. L. Leevy and T. M. Khoshgoftaar, "A survey and analysis of intrusion detection models based on cse-cic-ids2018 big data," *Journal of Big Data*, vol. 7, no. 1, pp. 1–19, 2020.
[26] M. R. Parsaei, M. J. Sobouti, S. R. Khayami and R. Javidan, "Network traffic classification using machine learning techniques over software defined networks," *International Journal of Advanced Computer Science Applications*, vol. 8, no. 7, pp. 220–225, 2017.
[27] J. Pei, Y. Chen and W. Ji, "A ddos attack detection method based on machine learning," *Journal of Physics: Conference Series*, vol. 1237, no. 3, 032040, 2019.
[28] H. Polat, O. Polat and A. Cetin, "Detecting ddos attacks in software-defined networks through feature selection methods and machine learning models," *Sustainability*, vol. 12, no. 3, 1035, 2020.
[29] Q. Li, L. Meng, Y. Zhang and J. Yan, "Ddos attacks detection using machine learning algorithms," in *Int. Forum on Digital TV and Wireless Multimedia Communications*. Singapore: Springer, pp. 205–216, 2019.
[30] M. M. Fadel, M. Sally, A. M. Ali-Eldin, M. K. Hassan and A. I. El-Desoky, "Hdlidp: A hybrid deep learning intrusion detection and prevention framework," *Computers, Materials & Continua*, vol. 73, no. 2, pp. 2293–2312, 2022.
[31] M. M. Fadel, S. M. El-Ghamrawy, A. M. Ali-Eldin, M. K. Hassan and A. I. El-Desoky, "The proposed hybrid deep learning intrusion prediction IOT (hdlip-IOT) framework," *PLoS One*, vol. 17, no. 7, e0271436, 2022.
[32] M. Rusyaidi, S. Jaf and Z. Ibrahim, "Machine learning method in detecting a distributed of service (ddos): A systematic literature review," *AIP Conference Proceedings*, vol. 2643, no. 1, 040034, 2023.
[33] W. Jiang, "Graph-based deep learning for communication networks: A survey," *Computer Communications*, vol. 185, pp. 40–54, 2022.
[34] Y. Li, R. Li, Z. Zhou, J. Guo, W. Yang *et al.,* "GraphDDoS: Effective DDoS attack detection using graph neural networks," in *2022 IEEE 25th Int. Conf. on Computer Supported Cooperative Work in Design (CSCWD)*, Hangzhou, China, pp. 1275–1280, 2022.
[35] Y. Cao, H. Jiang, Y. Deng, J. Wu, P. Zhou *et al.,* "Detecting and mitigating DDoS attacks in SDN using spatial-temporal graph convolutional network," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 6, pp. 3855–3872, 2021.

[36] L. Prehn, F. Lichtblau, C. Dietzel and A. Feldmann, "Peering only? Analyzing the reachability benefits of joining large ixps today," in *Int. Conf. on Passive and Active Network Measurement*, Cham, Switzerland, pp. 338–366, 2022.
[37] D. Bzdok, M. Krzywinski and N. Altman, "Machine learning: A primer," *Nature Methods*, vol. 14, no. 12, 1119, 2017.
[38] Y. Zhai, W. Song, X. Liu, L. Liu and X. Zhao, "A chi-square statistics based feature selection method in text classification," in *2018 IEEE 9th Int. Conf. on Software Engineering and Service Science (ICSESS)*, Beijing, China, pp. 160–163, 2018.
[39] D. Stiawan, M. Y. B. Idris, A. M. Bamhdi and R. Budiarto, "Cicids-2017 dataset feature analysis with information gain for anomaly detection," *IEEE Access*, vol. 8, pp. 132911–132921, 2020.