International Journal of Engineering Trends and Technology
ISSN: 2231-5381/ https://doi.org/10.14445/22315381/IJETT-V74I5P129

Volume 74 Issue 5, 458-483, May 2026

Original Article

Feature Selection based on Mutual Information and Machine Learning for DDoS Attacks Detection

MAZIGHI Abdellah[1], Lahoucine BALLIHI[2], Ghizlane ORHANOU[3]

[1,2]*LRIT Faculty of Sciences - Rabat, Mohammed V University in Rabat, Rabat, Morocco*
[3]*LabMIA-SI - Faculty of Sciences - Rabat, Mohammed V University in Rabat, Rabat, Morocco*

[1]*Corresponding Author : abdellah_mazighi@um5.ac.ma*

Received: 5 December 2025 Revised: 02 March 2026 Accepted: 28 March 2026 Published: 30 May 2026

Abstract *- Because of the dizzying increase of Distributed Denial of Service attacks (DDoS) all over the world and despite all the progress made in the field of the development of Intrusion Detection Systems (IDSs), there are still advances to be made in this area, particularly through the use of machine learning techniques. In the present paper, our main objective is to improve DDoS attacks detection by the use of Machine Learning techniques combined with feature selection based on Mutual Information. After the pre-processing step, we have proved by experiments on a recent large public dataset the positive effects of feature selection with Mutual Information on DDoS attacks detection performances. (Complexity, Resource consumption, Execution times and incorrectly classified). We dealt with high dimensionality of the dataset by feature selection with Mutual Information. Performance is evaluated by the use of relevant metrics such as accuracy, precision, recall, and F1-score. Finally, we conclude by analyzing our experimental results and propose some future works.*

Keywords *- CICDDoS-2019, DDoS attack, Intrusion detection, DDos detection, Machine Learning, Feature selection, Mutual Information.*

1. Introduction

People and institutions communicate every day with each other over the Internet. The number of Internet users is steadily growing, and it's expected to reach more than 7 billion users by 2029. Consequently, Internet infrastructure is targeted by attackers that try to compromise services in order to make profit [1]. The computer and Internet industry are the most frequently targeted and Denial-of-Service (DoS) attacks have increased considerably [2-4]. The main object of DoS attackers is to deprive a legitimate user of access to his resource. A DoS attack consists in a very large number of requests sent to the target server by the attackers. This large volume of requests overloads the victim's bandwidth and makes it unavailable to legitimate users. DoS attacks can be classified depending on target and behaviour [5]:

- Traffic-based attacks occur when an attacker overloads a victim's machine with TCP or UDP packets with the aim of lowering its performances.
- Bandwidth DoS attacks are the results of attackers sending huge quantities of anonymous data to a target machine in order to provoke a congestion of its bandwidth.
- Application-based DoS attacks are conceived and used to attack a particular system and they are not easy to mitigate [6, 7].

Moreover, these attacks have many other harmful aspects that damage institutions and their infrastructure. We can cite their increase in volume and frequency [7], and their devastating effects [8]. In fact, it is often difficult to know their origins and to find the real attackers. Witch is a constant and dangerous threat to organizations and their network infrastructure. As a response to all the mentioned threats, machine learning based intrusion detection systems (IDSs) are often used for the classification of the traffic to identify anomalies [9, 10]. IDSs are an important theme in cyber-security. Their role is to detect, classify, and alert intrusions in order to preserve availability, integrity, and confidentiality of a system [11]. There are three categories of intrusion detection systems: Signature-based intrusion detection systems (SIDS), Anomaly-based intrusion detection systems (AIDS) [9, 12, 13] and Hybrid Intrusion Detection Systems (HIDS):

- SIDSs: They are based on matching methods to detect previously known attacks with corresponding collected signatures. Known DDoS attacks are used to constitute the learning base. Their signatures are used to distinguish between attacks and normal traffic [7, 14-17].
- AIDSs: The main difference between SIDSs and AIDSs is that AIDSs can detect new attacks while SIDSs can't detect them. There are two phases in an AIDSs development project: a learning phase and a testing phase.

The learning phase is devoted to the use of normal traffic by the system to learn the normal behavior. During the testing phase, sets of features are used to check the ability of the system to detect attacks not seen before [9, 13, 18, 19, 20, 21].

- Hybrid IDSs: A Hybrid IDS is a combination of different intrusion detection systems with the aim of improving their effectiveness. Research has shown that combined algorithms perform better than single algorithms [22, 23].

The Machine learning-based IDSs perform better than static ones in terms of error rate, performance and responses to cyber-attacks [24, 25]. But it is to be noticed that machine learning-based IDSs offer better results only if they are trained on diverse, massive, and real-time datasets [26, 27]. There's a need for a large amount of harmful and harmless traffic data for training and testing steps [28].

In this paper, we have combined the use of machine learning techniques and feature selection with Mutual Information applied on the CICDDoS2019 dataset with the aim of improving DDoS attacks detection. After reviewing some relevant research works in the area of Machine Learning techniques for intrusion detection in section 2, we present, in section 3, a detailed description of Distributed Deny of Service attacks, their categories, their objectives, their operating modes and their classification in two categories (Exploitation attacks or Reflection attacks) depending on the methods applied by attackers. In Section 4, we present the technical background of machine learning algorithms that we have used to perform intrusion detection. We also present the metrics used to evaluate the performance of machine learning algorithms in terms of accuracy, precision, and time consuming. In section 5, we describe our methodology. It contains a brief description of the CICDDoS2019 dataset, the dataset preparation, sampling and cleaning. We have also presented in this section Mutual Information as a method used to select relevant features for traffic classification especially for a dataset containing a huge volume of traffic and a large number of features. In section 6, we present our approach including the experimental scheme and the five main steps. Section 7 contains the experimental results and their analysis. Finally, we present our conclusions with some potential future works in Section 8.

2. Related works

Many research articles related to the detection and prevention of DDoS attacks using machine learning techniques have been published. These articles cover different attack types, challenge, countermeasure techniques and other specific aspects of the phenomenon. In this section, we present some of the recent related works. Tamara Zhukabayeva et al. in [29] have presented an exhaustive study about the use of machine learning algorithms (RF and XGB) to detect botnet attacks on an IoT environment. They utilized the dataset N-BaIoT. It's collected from IoT commercial devices and it contains real traffic data including benign traffic and attacks used to identify Mirai and Bashlite. Their work showed that XGBoost and Random Forest outperformed other algorithms with accuracies between 93% and 98.92%.

Abdussalam Ahmed Alashhab et al. [30] proposed a traffic collector model Online Machine Learning (OML) based IDS and OML-based IPS. The proposed approach presents a real-time solution for detecting and mitigating DDoS attacks in SDN-based networks. They used Principal Component Analysis (PCA) as a feature selection methodology to improve accuracy and efficiency over CIC-DDoS2019 and InSDN datasets.

Afolabi and Akinola in [31] developed an IDS called KOMIG (knapsack optimization and mutual information gain). This IDS allies knapsack optimization, MI and machine learning techniques for better intrusion detection performances. They performed two (2) features selection methods, namely Knapsack Optimization and Mutual Information Gain, on the UNSW-NB15 dataset. Then they trained and tested four (4) machine learning algorithms (LR, RF, DT, and KNNs) on the optimized subset. The obtained results with KNNs-based KOMIG were the best and they vary between 97.14% and 99.46% for accuracy and recall respectively.

In [32], Alshhab et al. proposed an ensemble-based model that combines three classifiers (SGD, EBM, and MLP) for effective DDoS attack detection and mitigation in Software-Defined Networking (SDNs). They used Principal Component Analysis (PCA) as a feature selection methodology to improve accuracy and efficiency using CIC-DDoS2019 and InSDN datasets.

Ahmed Mohamed Salama et al. in [33] presented a study about the use of machine learning algorithms to detect and lessen the effects of DDoS attacks. The used algorithms, namely KNN, RT, NB, Stochastic Gradient Boosting, Logistic Regression, and SVM were trained and evaluated using the CICDDoS2019 dataset. The obtained results showed that DDoS attacks were perfectly detected with all the used algorithms but SVM outclassed them with an accuracy of 99 %.

Yongqiang Shang [34] presented a machine learning based approach to spotting distributed denial of service (DDoS) attacks against servers situated in the cloud. He used Nearest Neighbor, Random Forest, and Naive Bayes to detect a distributed denial of service attack with a 99.75% detection rate.

Alamgir Hossain and Saiful Islam [35] proposed a hybrid feature selection method based on Correlation Analysis (CA), Mutual Information (MI), and Principal Component Analysis (PCA) with an ensemble learning based machine learning classifier with aiming to enhance DDoS attacks detection. They trained and tested their model on CIC-DDoS2019,

DDoS-SDN, CSE–CIC–IDS2018, APA-DDoS, and DDoS-botnet datasets. The conclusion was that the use of the hybrid feature selection method and an ensemble-based machine learning classifier leads to excellent DDoS attacks detection performances.

Hakem Beitollahi et al. [36], were interested in application layer Distributed Denial of Service (App-DDoS) attacks which continue to be a pervasive problem in cybersecurity. They presented a highly effective and adaptable solution for detecting various types of App-DDoS attacks. They combined Random Forest (RF), Gaussian Mixture Models (GMM) and a human with expertise in DDoS to enhance the detection of those attacks. Using various machine learning algorithms (GGM, DT, GA etc.) with feature selection strategy and combining CICIDS2017 and CICIDS2019 datasets, they minimized false alarms and maximized accuracy, precision, recall, and F1 score.

Siriporn Chimphlee and Witcha Chimphlee [37] proposed a machine learning based classification scheme for IDS. They used the CSE-CIC-IDS-2018 dataset after performing its pre-processing techniques (under-sampling, feature selection). Then, they used some classifier algorithms to choose the best performing model to classify attacks. They have implemented and compared seven algorithms (Random Forest (RF, Linear Regression (LR), K-Nearest Neighbors (KNN), classification and regression trees (CART), Bayes, RF, multilayer perceptron (MLP), XGBoost) and used various criteria in order to implement IDSs.

Mohammad Najafimehr et al. [38] have highlighted the dangers of DDoS attacks and the importance of developing effective defense mechanisms. They provided an analysis of some machine learning approaches and presented some methods based on machine learning to detect DDos attacks. As a conclusion, they stated that combining supervised and unsupervised methods along with non-ML methods may be the best approach to detect known or unknown attacks. They have also presented a survey and taxonomy of DDoS attacks and machine learning based detection methods. This research article also presents a comprehensive taxonomy of machine learning based DDoS detection methods such as supervised, unsupervised, hybrid approaches and a review of their challenges.

Seyedakbar Mostafavi et al. [39] have also presented a survey and a taxonomy of DDoS attacks and machine learning based detection methods. They presented the importance and the challenges in developing defense mechanisms against DDoS attacks given their diverse types, networks heterogeneity and complexity of the communication protocols. Furthermore, the appearance of the reflective DDoS attacks (DrDoS) constitutes a major threat to existing countermeasures. This research article also presents a comprehensive taxonomy of machine learning based DDoS detection methods such as supervised, unsupervised, hybrid approaches and a review of their challenges.

In [40], Erick Odhiambo Omuya et al. developed a sentiment analysis model based on machine learning and used data where dimensionality reduction is incorporated. Their model was tested with NB, SVM, and KNNs algorithms. The results from experiments done on the proposed model showed that training the model on preprocessed data and reducing dimensions greatly improve the performances of the model.
Alduailij et al. in [5], presented a method based on machine learning with Mutual Information (MI) and Random Forest Feature Importance (RFFI) to detect DDoS attacks in the cloud computing. They experienced five (5) machine learning algorithms, namely: RF, Gradient Boosting (GB), Weighted Voting Ensemble (WVE), KNNs, and LR on the dataset constituted with 19 selected features. Accuracies of 0.99 were obtained with RF, GB, WVE, and KNN.

Lakshmeeswari and Shadi [41] developed a machine learning model called SWASTHIKA which uses feature transformation traffic for DDoS attacks detection in Industry 4.0. They also presented a new Gaussian based traffic attribute pattern similarity function for evolutionary feature clustering to achieve feature transformation-based dimensionality reduction and a Gaussian based network traffic similarity function for similarity computation between network traffic instances.

Elijah M. Maseno et al. [22] reviewed more than 100 studies related to hybrid intrusion detection systems between 2012 and 2022. They noticed that there were gaps in the development of hybrid intrusion detection systems and a need of further research in the area.

Odhiambo Omuya et al. [42] proposed a hybrid model called Principal Component Analysis and Information Gain (PCA-IG) for feature selection to support classification and compared it to some other algorithms like Naive Bayes or J48. They used breast cancer data set for their experiments and concluded that PCA-IG model accuracy and general performance were performed after feature selection.

Dhinsa et al. in [43], have automated identification and classification of microbes by the use of five (5) machine learning algorithms namely Multiple Layer Perceptron (MLP), K-Nearest Neighbors (KNN), Logistic Regression (LR), Support Vector Machine (SVM) and quadratic discriminant analysis (QDA). These algorithms were combined with Mutual Information (MI) and Principal Component Analysis (PCA) as feature selection methods. They have produced their own dataset by collecting thousands of microbe slides and microscopic images. They have proved by experiment that Mutual Information is the best method for feature selection and SVM outperforms other machine learning algorithms.

Md Al-Imran and Shamim H. Ripon [44] presented an intrusion detection system containing three phases: The first experiments have been conducted with SVM, Decision Tree, and KNN. The second ones were conducted applying Random Forest and XGBoost as they usually show significant performance improvement in supervised learning. The third experiments were conducted using deep learning techniques. They used Feed Forward, LSTM, and Gated Recurrent Unit neural network algorithms and Kyoto Honeypot Dataset for their experimental purpose. The results showed a significant improvement in IDS and applicability of the proposed model in IDSs.

Mohammed Al-Sarem et al. in [45], proposed an aggregated method of feature selection based on Mutual Information combined with machine learning techniques with the objective of enhancing the detection of Botnet attacks in an IoT environment. They used the N-BaIoT dataset and performed an aggregated feature selection method on it. The feature selection methods MI, PCA and ANOVA f-test were combined to select the most relevant features. After that, six (6) machine learning algorithms (RF, XGB, Gaussian Naive Bayes, KNNs, LR, and SVM) were trained and tested on the N-BaIoT dataset. The obtained results revealed that XGB and KNN outperformed other classifiers with accuracies of 99.19% and 98.28% respectively. Majid Torabi et al. [46] developed a review of on Feature Selection and Ensemble Techniques for Intrusion Detection System (IDS). They have made a classification of detection methods in IDS, a Classification of machine learning techniques used in anomaly-based IDS, an identification of feature selection for anomaly-based IDS and an identification of ensemble classification for anomaly-based IDS.

Sharafaldin et al. [47] have reviewed and analysed the main existing DDoS attack datasets such as CAIDA UCSD or DARPA 2000 aiming to find their limitations and shortcomings. They took those limits and shortcomings into account and addressed them while generating their new dataset named CICDDoS2019. The CICFlowMeter software was used to extract and calculate the 80 network traffic features. They have proposed a new taxonomy for DDos attacks and produced a new dataset dedicated to them. In addition, crucial attributes for every type of attack are provided using Random Forest Regressor (RFR).

A.Ghorbani et al. in [48], have proposed a new taxonomy for DDos attacks and produced a new dataset dedicated to them. In addition, crucial attributes for every type of attack are provided using the random forest regressor (RFR). Various machine learning algorithms such as Naive Bayes (NB), Random Forest (RF), Logistic Regression (LR), and ID3 were implemented on this dataset. The study concluded that ID3 performed well. In order to find the most important features, they used the feature selection technique to detect different attacks including Distributed reflective Deny of Service (DrDoS).

Tasnuva Mahjabin et al. [7] made a comprehensive survey about distributed denial-of-service attack, prevention, and mitigation techniques. They presented a detailed and systematic analysis of those attacks, their motivations and evolution, and protection and mitigation techniques. They also presented some statistics such as the volume sizes of DDoS attacks worldwide and the numbers and proportions of attacked sites by sector.

Pavle Vuletic Ognjen Joldzic et al. [19] presented a solution for detecting and mitigating Denial-of-Service (DoS) attacks. It's a distributed scalable solution for the detection of lower-level Denial-of-Service (DoS) attacks which are among the most common types of attacks causing the disruption of service operations and reliability. The system monitors network traffic for denial-of-service attacks and is capable of preventing some of them.

3. Denial of Service Attacks (DoS)

There are three main objectives in computer security: Availability, Confidentiality, and Integrity. Availability is defined as the ability to use the desired information or resource in a reliable and timely manner. Denial of Service (DDoS) attacks aim to disrupt websites and online services for users, making them unavailable by overloading them with much more traffic than they can handle.

There are two types of DDoS attacks regarding the method used by attackers: reflection attacks, where the attacker sends a huge number of packets to saturate the target machine and exploitation attacks, where the attacker exploits a vulnerability in the target machine to make it out of service. [49-53]. There are three main types of DDoS attacks:

3.1. Simple Denial of Service attacks

Denial of Service is a threat that can violate the availability of a resource in a system. On the other hand, a Denial-of-Service Attack, is an action (or a set of actions) executed by a malicious entity to make a resource unavailable to its users [49-51]. In DoS attack, the attacker initiates transmission of malicious traffic through the handlers and a set of compromised machines to make a set of target machines inaccessible. Network security companies affirm that DoS attacks are one of the greatest concerns for service providers. [50].

3.2. Distributed Denial of Service Attacks (DDoS)

DoS attack, also called non-distributed DoS attacks, methods and tools are becoming more sophisticated, effective, and also more difficult to trace to the real attackers [54]. But DDoS attacks are more difficult to mitigate and detect. In a typical DDoS attack, a large number of compromised hosts are

grouped to send useless and harmful packets to jam a victim or its Internet connection, or both.

They can be classified as internal DDoS attacks and external DDoS attacks [55], but both are similar in their execution and their effects. There's an important difference between distributed and non-distributed DoS attacks from the standpoint of the security device that is supposed to mitigate and detect them [19].

3.2.1. DDoS Attacks Taxonomy

The researchers have proposed various taxonomies and classification of DoS attacks [56, 57]. For example, Kotapati et al. divided the attacks into interception, fabrication, modification, interruption and denial of service [58]. As illustrated in Figure 1, attacks on the CICDDoS2019 dataset were carried out using TCP/UDP based protocols of the application layer. Figure 1 shows the detailed taxonomy of DDoS attacks and splits them into reflection-based and exploitation-based attacks [59].

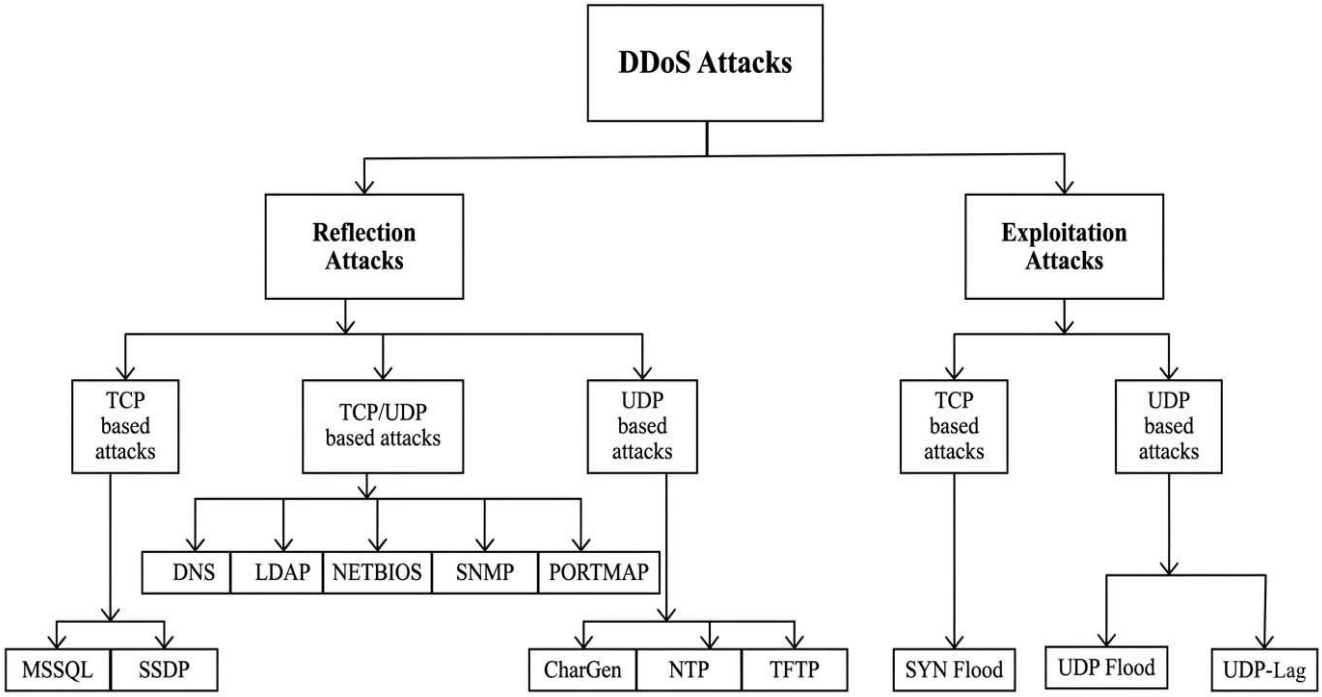


Fig. 1 DDoS attacks taxonomy

3.2.2. Reflection based DDoS Attacks (DrDDoS)

The method applied by the DDoS attackers is the reflection method, where compromised servers are used to forward traffic to the victim machines and assist in consuming the victims' resources. This method helps the attacker to send traffic to the victim indirectly and that helps him to remain unseen and undetected. In this method, all attack packets sent by the attacker contain the IP address of the victim in the source address field of IP packets. When a server receives these service requests, it sends its response to the victim node, not to the actual source node of packet, in which, the attacker controls the master and slave zombies and instructs them to flood the request packets to the reflector node to bring down the victim [55, 59, 60]. Reflection based attacks are carried out through application layer protocols (Transmission Control Protocol (TCP) and User Datagram Protocol (UDP) or a combination of them). As illustrated in Figure1, TCP based attacks include MSSQL and UDP. UDP based attacks include CharGen, NTP and TFTP. TCP/UDP based attacks include DNS, LDAP, NETBIOS, SNMP and PORTMAP.

3.2.3. Exploitation based DDoS Attacks

Application layer protocols such as TCP and UDP are used to carry out these attacks. TCP based exploitation attacks include SYN Flood while UDP based exploitation attacks include UDP-Flood and UDP-Lag. An UDP flood attack is executed by sending a large number of UDP packets. They are sent to random ports at a very high rate aiming to exhaust the bandwidth of the network and degrading the system performances. TCP based SYN flood also degrades dramatically server resources. SYN flood is executed by sending repeated SYN packets to the victim machine until the server crashes. The UDP-Lag attack disrupts the connection between the client and the server. It's mostly used in gaming where the players try to slow down or interrupt the movements of each other.

4. Machine Learning for Anomaly Detection

Machine learning could be defined as the process of extracting knowledge from a large quantity of data. Machine learning models comprise rules, methods and complex

algorithms applied to extract interesting knowledge, to recognize or predict a behavior. Several machine learning algorithms have been applied especially in the area of AIDSs.

Techniques like clustering, Nearest Neighbors, association rules, Decision Trees, Neural Networks and Genetic Algorithms have been used. ML algorithms may be compared taking in consideration some parameters like accuracy, speed, scalability and learning ability.

For our experiments, we have used four ML algorithms, and compared them according to some evaluation criteria usually used for the model's performance evaluations. [13, 15, 61-66].

4.1. Machine Learning Algorithms

Decision tree (DT): A decision tree is a nonparametric supervised learning algorithm, which is utilized for both classification and regression tasks. It's built on a collection of attributes (features). A root node represents a test attribute. Branches represent the results. Decision leaves show the final decision taken after calculating all features in the form of a class label. [67, 68].

Extra Trees (ET): Also called Extremely Randomized Trees is an ensemble learning method. It consists of implementing extra randomness when growing trees on sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. As a result, it trades more bias for a lower variance.

Random Forest (RF): This decision tree-based classifier was proposed by Breiman in 2001. It's based on the principle that is strong learner group is created by a group of weak learners. It's also one of the ensemble classifier methods. It combines multiple classifiers to produce a hypothesis of a problem to set up a typical result.

Each decision tree is created through a random selection of attributes at each node for separation. The final result of a classification can be decided by majority voting or weighted voting. Random Forest has an interesting advantage that is its variance decreases as the number of trees increases while the bias remains the same. [65, 69, 70].

XGBoost or Extreme Gradient Boosting (XGB): It' s a supervised learning algorithm whose principle is to combine the results of a set of simpler and weaker models in order to provide a better prediction. It' s a scalable and distributed gradient boosted decision tree machine learning library.

The principle is that instead of using a single model, the algorithm will use several, which will then be combined to obtain a single result. It works sequentially and provides parallel tree boosting. XGB is the leading machine learning library for regression, classification, and ranking problems.

4.2. Performance Metrics

IDSs are typically evaluated based on many evaluation metrics. The evaluation results are usually recapitulated in a confusion matrix. We have used the standard performance measures defined below in order to calculate the evaluation metrics [71].

- True Positives (TP): TP, also called correct detection, is the number of attacks correctly classified as attacks by the model.
- False Positives (FP): FP, also called Type-1 Error, is the number of the benign data wrongly classified as attack by the model.
- False Negatives (FN): FN, also called Type-2 Error, is the number of attacks wrongly classified as benign or normal traffic by the model.
- True Negatives (TN): TN, also called Correct Rejection is the number of benign data correctly classified as benign by the model.

We have used the above-mentioned metrics for the performance evaluation and comparison of the classifier algorithms (DT, ET, RF and XGB) before and after Feature Selection with Mutual Information:

Accuracy: Measures the overall correctness of the classifier by calculating the ratio of correctly classified instances (both intrusions and normal traffic) to the total number of instances. Accuracy is calculated using the formula (1) below:

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

Precision: Also known as positive predictive value, precision signifies the proportion of correctly detected intrusions among all instances classified as intrusions. It evaluates the system' s ability to avoid false positives. Precision is calculated using the formula (2) below:

$$\text{Precision} = \frac{TP}{FP + TP} \quad (2)$$

Recall or True Positive Rate (TPR): Recall, also known as the true positive rate (TPR), is the percentage of normal instances correctly classified as normal traffic by the system out of the total of normal instances. It' s calculated using the formula (3) below:

$$\text{Recall} = \text{TP Rate} = \frac{TP}{TP + FN} \quad (3)$$

F1-Score: Also known as balanced F-score or F-measure, it can be interpreted as a harmonic mean of the precision and recall. The formula for the F1- score is given in (4):

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

4.3. Analysis Tools

We have executed our experiments on a computer with the specifications below: Processor: Intel Core i7, CPU: 1.80GHz, RAM: 32 Go, hard disk capacity: 256 Go SSD, system type: 64 bits, graphics card: 32 Gb. We have also used Python and Pandas library for data manipulation and analysis.

5. Pre-Processing the CICDDoS2019 Dataset

The performances of a machine learning algorithm are highly dependent on the datasets used for training and testing the algorithm. We have used the CICDDoS2019 dataset to train and test machine learning models to predict DDoS attacks.

5.1. Brief Description of the The CICDDoS 2019 Dataset

The CICDDoS2019 dataset is a public dataset from the Canadian Institute of Cybersecurity (CIC) at the University of New Brunswick (UNB). It is a labeled public dataset. It's a recent dataset and it contains 13 types of Distributed and reflected Denial of Service attacks (DrDoS).

It also contains more than 70 million labeled instances (Normal traffic and attacks) and 88 features extracted using CICFlowMeter[72].

Attacks in this dataset are classified into two categories: Reflection-based and exploitation-based attacks. These attacks perform through diverse protocols like Transmission Control Protocol (TCP), User Datagram Protocol (UDP) or a combination of both. They include MSSQL, SSDP, DNS, LDAP, NetBIOS, SNMP, PortMap, CharGen, NTP, and TFTP [60, 73].

It contains two main families of traffic. Normal traffic or benign attacks and malicious traffic mainly DDoS attacks. The dataset consists of 12 CSV files related to DDoS attacks executed on training day and 7 CSV files related to DDoS attacks executed on testing day.

This dataset is publicly available at: http://www.unb.ca/cic/datasets/CICDDoS2019 [74, 75].

5.2. CICDDoS2019 Dataset Preparation and Cleaning

Preparing and cleaning a dataset with the aim of making it ready for the use in a machine learning project usually takes a lot of time. Some researchers claim that this operation is the most important of the project and that it may take more than 80 % of the total time of the project [64, 76, 77].

The CICDDoS2019 dataset originally contains roughly 70,427,863 instances divided in 12 CSV files related to DDoS attacks executed on training day and 7 CSV files related to attacks executed on testing day.

Training day attacks include NTP, DNS, LDAP, MSSQL, NetBIOS, SNMP, SSDP, UDP, UDPLag, WebDDoS, SYN and TFTP. Testing day attacks are Portscan, NetBIOS, LDAP, MSSQL, UDP, UDPLag and SYN. Each file originally contains 88 features.

But, we should notice that high dimensional data with irrelevant and redundant features and imbalance usually present big challenge to data analysis algorithms.

That's why the pre-processing step of the CICDDoS2019 dataset will constitute a very important part of this work [37, 40, 42, 47].

Preparing and cleaning data includes many steps and operations. The CICDDoS2019 dataset contains some attributes that are useless for the machine learning process.

They won't have any positive impact on the performance of the models. We have removed the attribute called 'Fwd Header Length.1' because it's a redundant one.

Most public datasets used in cybersecurity contain a lot of duplicated records [78]. We have removed duplicated instances because they are worthless for training and testing the models.

For several reasons, there is always a lot of missing values in a dataset. We have dealt with this problem by using statistical inference [80].

The raw original csv files contain a lot of useless information such as infinite values and NaN values. They need to be cleaned. Two attributes namely, 'Flow Bytes/s' and 'Flow Packets/s' contain infinite values. 'Flow Bytes/s' contains 38,410 infinite values while 'Flow Packets/s' contains 43,432 infinite values.

Figure 2 illustrates the Pie diagram of attacks in CICDDoS2019 Subset after preprocessing. Now, after this operation, the 5% raw csv file sample contains 83 features and 3,520,214 instances.

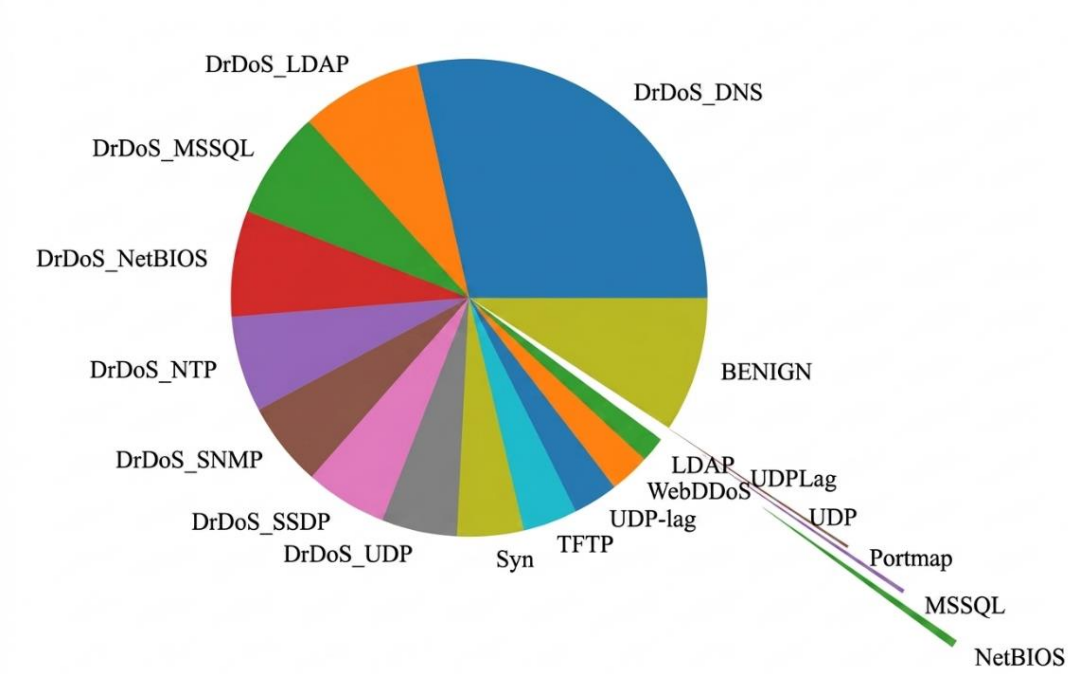


Fig. 2 Pie Diagram of attacks in the Subset

5.3. Relabeling

We have used a CICDDoS2019 subset containing only 5% of the original dataset but it still remains a huge CSV file because it contains 3,520,214 instances. This subset includes a class label and it is characterized by high class imbalance which may affect detection performances. To deal with high class imbalance, we have proposed three relabeling schemes: Binary Classification, Three Classes Classification and All Attacks Classification. Relabeling details are given in Table 1.

5.4. Cleaned and Relabeled Dataset

After the pre-processing phase, we obtain cleaned and ready to be exploited files. The three obtained files constitute the CICDDoS2019 subsets (Binary Classification, Three Classes Classification, and All Attacks Classification).

5.5. Feature Selection

5.5.1. Advantages of Feature Selection

Dimensionality is one of the most important issues in large datasets. The feature selection process reduces dimensionality by selecting the most relevant features. It has many advantages, such as dimensionality reduction, training machine learning algorithm rapidly, models complexity reduction, improving accuracy, and Over-fitting avoiding.

5.5.2. Feature Selection Methods

There are different methods for feature selection: filter methods, (Chi-Square Test, ANOVA F-test, Mutual Information) wrapper and embedded methods (Recursive Feature Elimination, Sequential Feature Selection, L1-Regularization, Tree-Based Feature Importance).

5.5.3. The Selected Method

For our experiments, we have used a filter-based method for feature selection called Mutual Information (MI) because it's suitable for high dimensional datasets like CICDDoS2019 dataset. MI is a non-negative value which measures the mutual dependence between the two variables.

If the variables are independent, then their MI is equal to zero. Higher values of MI mean high dependence between the variables. The concept of Mutual Information is intimately linked to entropy of a random variable.

Entropy is a fundamental notion in information theory that quantifies the expected 'amount of information' held in a random variable. Its calculation formula between two clusters U and V is given as in equation (5) where |Ui| is the number of the samples in cluster Ui and |Vj | is the number of the samples in cluster Vj.

$$MI(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \frac{N|U_i \cap V_j|}{|U_i||V_j|} \qquad (5)$$

5.5.4. Ranking and Groups Constitution

The feature scores and ranks which are the results of feature selection by mutual information are tabulated in Table 3. After that, we have grouped the features according to their scores. Table 2 illustrates the six created groups which will constitute the subsets of our experiments. Table 10 illustrates the features ranking after feature selection with Mutual Information.

6. Proposed Approach

As shown in Figures 3 and 4, our experimental scheme includes 5 main tasks. They are detailed as follows:

At first, we have downloaded the 12 CSV files constituting the training bed and the 7 CSV files constituting the testing bed.

Then, we have sampled each file to 5%. Finally, we have merged all the 19 samples to obtain a unique 5% file sample that constitutes our raw data subset. Figure 6 illustrates the details of this operation.

For the second task, we have removed redundant features (Fw Header length 1) and useless features or those generated automatically after the merging operation (Unnamed 0.1,), treated NaN values and infinity values by applying statistical inference, encoded all object types, and normalized our data. Finally, we have proposed three relabeling schemes to analyze the subset (Binary Analysis, Three Classes Analysis and All Classes Analysis).

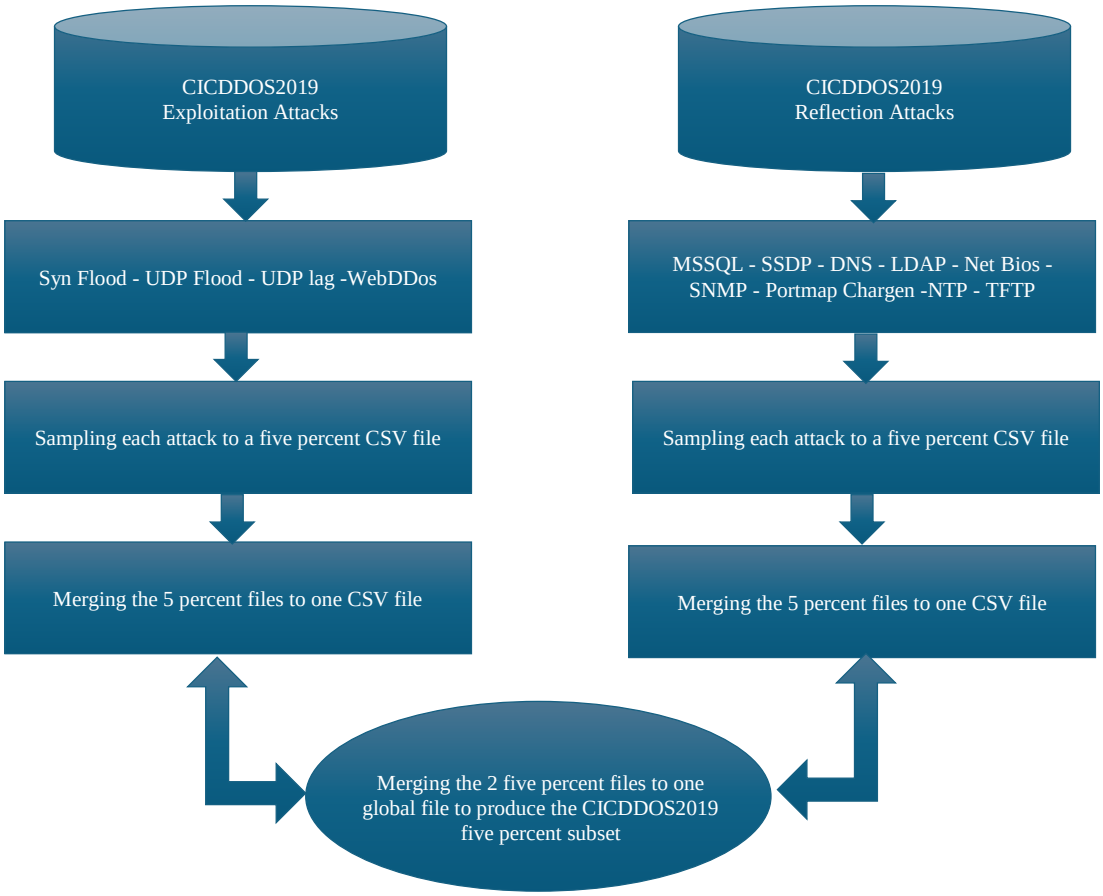


Fig. 3 CICDDoS2019 dataset sampling to a five per cent subset

Table 1. CICDDoS2019 Subset Relabeling Table

All Attacks Classification		Thress Classes Classification		Binary Classification	
Old Labels	**Instances**	**New Labels**	**Instances**	**New Labels**	**Instances**
Benign	5,577	Benign	5.577	Benign	5.577
Syn	322,546	Exploitation Attacks	534,344	Attack	3,514,637
UDP	193,369				
UDP-lag	18,330				
UDP-Lag	89				
TFTP	1,004,183	Reflection Attacks	2,980,303		
MSSQL	289,429				

DrDoS-SNMP	257,99				
DrDoS-DNS	253,551				
DrDoS-MSSQL	226,134				
DrDoS-NetBIOS	204,68				
NetBIOS	182,996				
DrDoS-UDP	156,739				
DrDoS-SSDP	130,541				
DrDoS-LDAP	108,98				
LDAP	95,613				
DrDoS-NTP	60,106				
Portmap	9,340				
WebDDoS	20				
Total All Attacks	3,520,214	Total 3 Classes	3,520,214	Total Binary Analysis	3,520,214

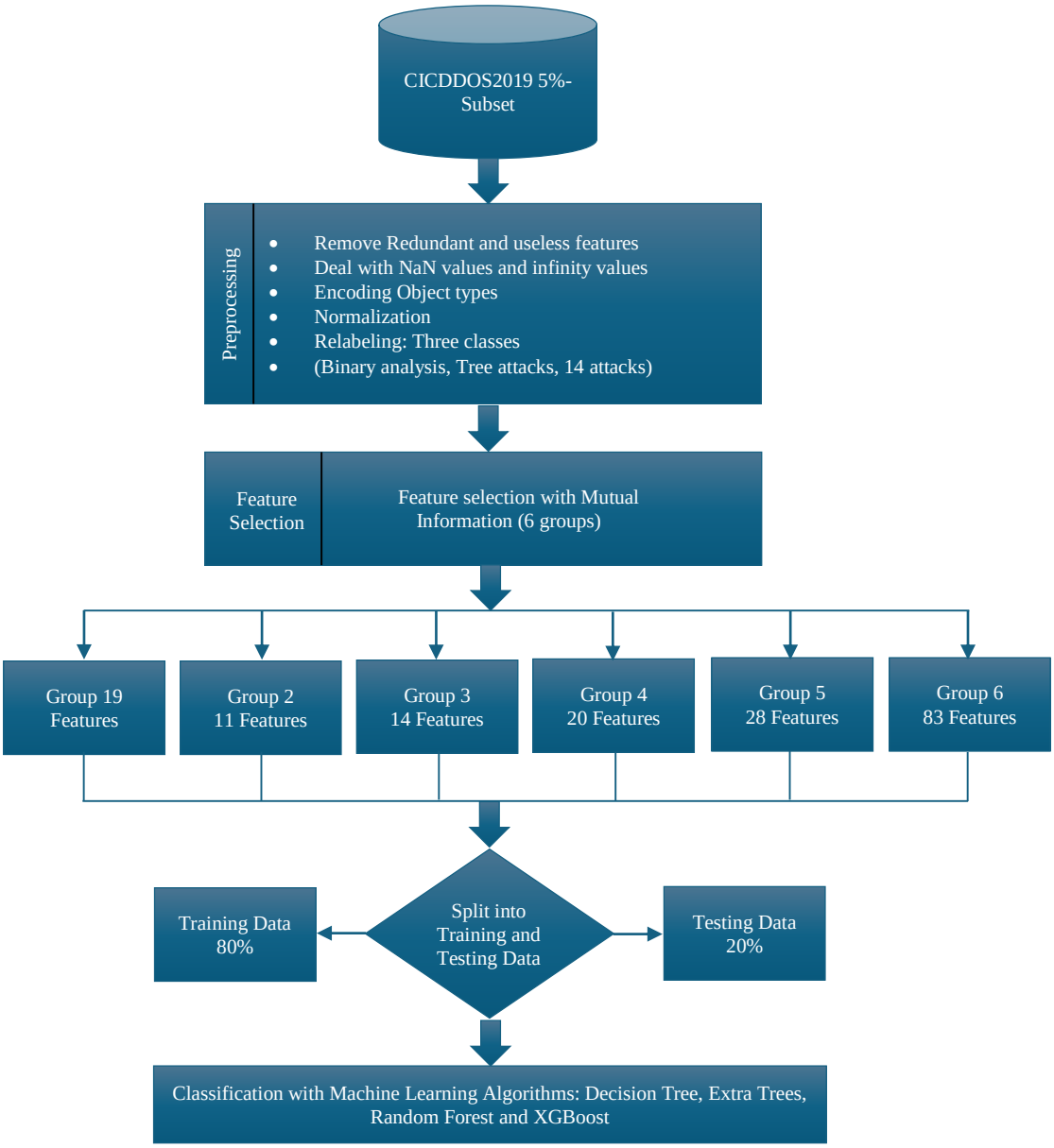
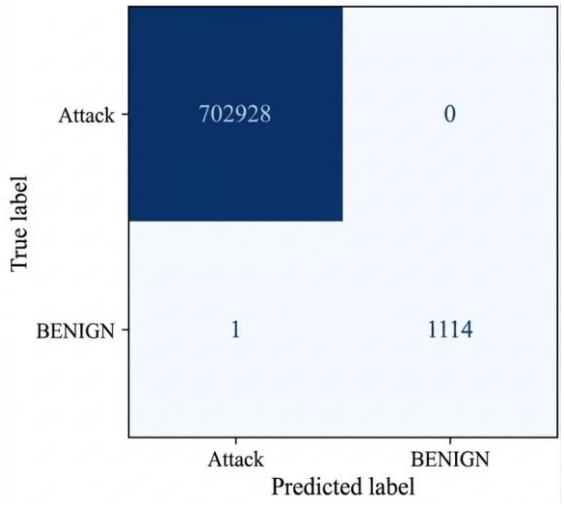


Fig. 4 Experimental Scheme

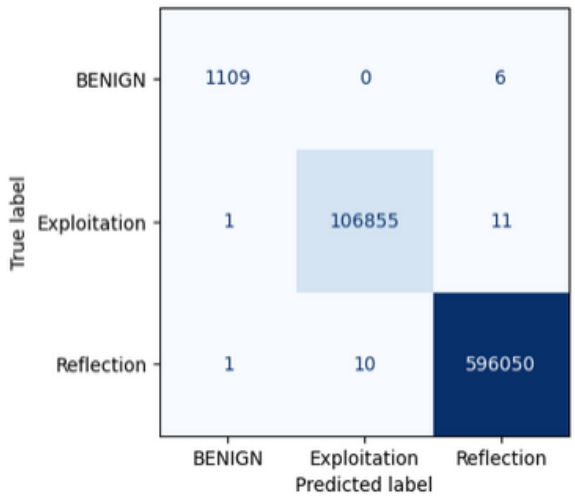
Table 2. Groups of Selected Features with Mutual Information

% of cumulated scores	Selected features	Groups	Feature Ranks
50%	9	Group 1	1 to 9
60%	11	Group 2	1 to 11
70%	14	Group 3	1 to 14
80%	20	Group 4	1 to 20
90%	28	Group 5	1 to 28
100%	83	Group 6	1 to 83

Next, we have performed feature selection with Mutual Information on all the 3 subsets generated after the relabeling operations (Binary Analysis, Three Classes, and All Classes). After calculation of Mutual Information scores, we have ranked the features according to their scores. Then we have divided the family into six groups according to their cumulative Mutual Information scores. Thus, we have created six groups of features. Table 2 shows the results of the operation.



(a) Binary Classification



(b) Three types of attacks

Fig. 5 Confusion Matrix obtained with DT Classifier for Binary Classification

Then, we have experienced four (4) Machine Learning algorithms (DT, ET, RF and XGB) on each group of features and labeling class. The results are given in the form of confusion matrices and classification reports for each group of features. Execution time, Accuracy, Precision, Recall, F1-Score and the number of incorrectly classified are the metrics we have used to evaluate our models. Table 8 illustrates a comparison between the numbers of incorrectly classified for Group 1, Group 4 and Group 6.

Lastly, we have compared and analyzed all the results and drew our conclusions. Figures 7a,7b and 9 represent examples of the generated confusion matrices.

7. Experimental Results

As illustrated in Figure 4, we have performed feature selection with Mutual Information and used four (4) machine learning classifier algorithms (DT, ET, RF and XGB). We have also created three subsets for our experiments and named them: Binary Classification, Three Classes Classification, and All Classes Classification. The details are shown in Table 1. The feature section process with Mutual Information led us to the creation of six groups of features. Group 1 contains only the nine (9) most relevant features, Group 2 contains eleven (11) features, Group 3 contains fourteen (14) features, Group 4 contains twenty (20) features, Group 5 contains 28 features and Group 6 contains all the features (83).

Table 2 gives the details about the six created groups of features. In order to evaluate the performance of Feature Selection with Mutual Information and the four (4) algorithms, we have used the most relevant evaluation metrics namely Execution Time, True Positives (TP), False positives (FP), True Negatives (TN), Precision, Recall, F1-Score, Accuracy, and Number of incorrectly classified. We have performed feature selection with Mutual Information on each one of the three subsets generated after the relabeling operation (Binary Classification, Three attacks Classification and All Attacks Classification). Then, we have experienced the four (4) classifier algorithms over all the subsets.

7.1. Execution Times

All the results in terms of execution times for all groups and classifications are given in Table 5. They are also visualized in Figure 13. Using Binary Classification XGB and DT outperform ET and RF for all groups. For the Three Classes analysis, XGB and DT outperform with all groups and

they realize nearly the same execution times for all groups. Using All Attacks Classification subset, DT outperforms all other classifiers in terms of execution times for all groups. XGB comes second, and ET the third. RF is the slowest among the four (4) classifiers. But, as a dilemma, it realizes better detection performances. The comparison of obtained results confirms that execution times are affected by the number of features selected by Mutual Information.

7.2. Other Global Performance Results Analysis

Group 1, nine (9) most relevant features: For the binary classification, the performances are tabulated in Table 6. Also, Figure 14 illustrates the numbers of incorrectly classified attacks. For the incorrectly classified, DT, ET and RF have nearly the same results but XGB produces more False Positives and more False Negatives than others. The results for Group 1 using the Three Attacks Classification are tabulated in Tables 5, 7, 9 and visualized in Figures 13 and 14. They show that DT and XGB are faster than ET and RF and have low execution times, but RF is the slowest with long execution times. RF, DT and ET perform well for detecting the three attacks but XGB has high False Positive Rates. This classifier produces a lot of FP and it has some lack of precision to distinguish between Exploitation and reflection attacks. When we consider the All Attacks Classification, DT, RF and ET surpass XGB with 374, 377, 464 and 5,391 total incorrectly classified attacks respectively. DT outperforms with its lowest execution time. Then comes XGB, ET and RF. DT detects correctly all attacks with precisions of more than 99.5% except Benign attacks with a precision of 90%, UDP-Lag with 99%, and WebDDoS with 43%. RF and ET detect correctly all attacks with precisions of more than 99.5 % except Benign attacks with a precision of 91%, UDP-Lag with 99%,UDPLag with 83% and WebDDoS with 40%. XGB detects DrDoSMSSQL, DrDoS-SSDP, LDAP, MSSQL, NetBIOS, TFTP, UDP and WebDDoS with precisions of more than 99.5%. But some attacks like UDPLag, Portmap or Benign attacks are detected with precisions of 69%, 70% and 84% respectively.

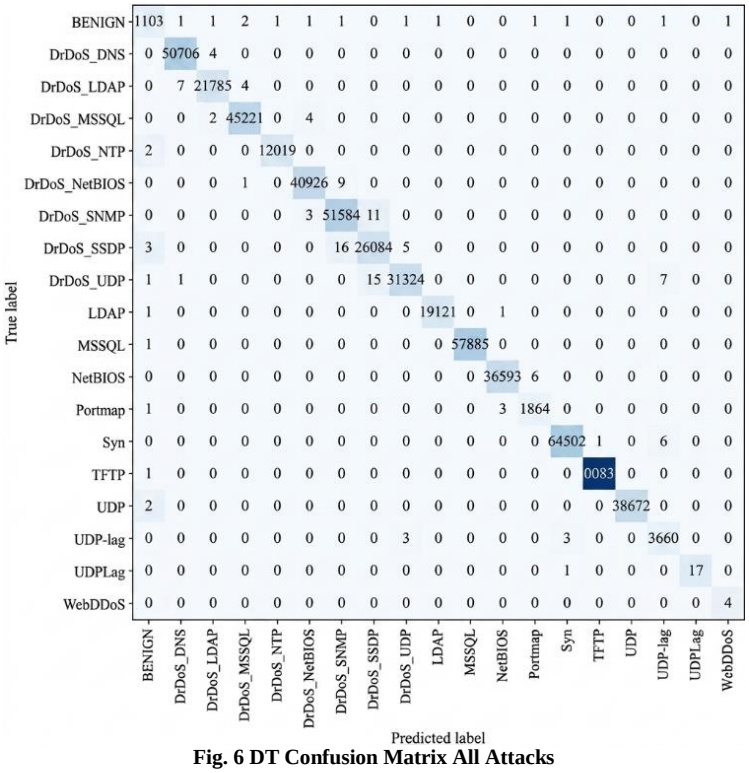


Fig. 6 DT Confusion Matrix All Attacks

Table 3. Feature Scores and Ranks with Mutual Information in the CICDDoS2019 dataset

Ranks	Feature Names	Scores	Ranks	Feature Names	Scores
1	Timestamp	2,4036	43	Subflow Bwd Packets	0,0833
2	Packet Length Mean	1,8134	44	Bwd Packet Length Mean	0,0788
3	Fwd Packet Length Mean	1,8093	45	Bwd Packet Length Min	0,0785
4	Fwd Packet Length Min	1,8092	46	Avg Bwd Segment Size	0,0785
5	Avg Fwd Segment Size	1,8092	47	Total Length of Bwd Packets	0,0777
6	Max Packet Length	1,8091	48	Subflow Bwd Bytes	0,0774
7	Min Packet Length	1,8087	49	Bwd Packet Length Max	0,0774

8	Fwd Packet Length Max	1,8079	50	Bwd IAT Max	0,0768
9	Average Packet Size	1,7771	51	Bwd IAT Total	0,0765
10	Subflow Fwd Bytes	1,7517	52	Bwd IAT Mean	0,0764
11	Total Length of Fwd Packets	1,7514	53	Bwd Header Length	0,0737
12	Flow Bytes/s	1,6906	54	Bwd IAT Min	0,0725
13	Source Port	1,004	55	Down/Up Ratio	0,0656
14	Fwd Packets/s	0,6411	56	Idle Min	0,026
15	Flow Packets/s	0,6141	57	Idle Max	0,0257
16	Flow IAT Mean	0,5994	58	Active Max	0,0256
17	Flow Duration	0,594	59	Idle Mean	0,0253
18	Flow IAT Max	0,5751	60	Active Mean	0,0249
19	Fwd IAT Mean	0,5492	61	Active Min	0,0247
20	Fwd IAT Total	0,5463	62	Idle Std	0,0233
21	Fwd IAT Max	0,5374	63	Destination Port	0,0194
22	Flow IAT Std	0,5145	64	Active Std	0,019
23	Fwd IAT Std	0,4591	65	Bwd IAT Std	0,0122
24	Protocol	0,3972	66	URG Flag Count	0,005
25	act-data-pkt-fwd	0,363	67	Fwd PSH Flags	0,0021
26	Init-Win-bytes-forward	0,3281	68	Bwd Packet Length Std	0,0021
27	ACK Flag Count	0,3194	69	RST Flag Count	0,002
28	Subflow Fwd Packets	0,2968	70	CWE Flag Count	0,0014
29	Total Fwd Packets	0,2965	71	Fwd Avg Packets/Bulk	0,0006
30	Packet Length Std	0,1878	72	Bwd Avg Bytes/Bulk	0,0005
31	Packet Length Variance	0,1871	73	PSH Flag Count	0,0003
32	Fwd Packet Length Std	0,1846	74	Fwd URG Flags	0,0003
33	min-seg-size-forward	0,149	75	Fwd Avg Bulk Rate	0,0003
34	Flow IAT Min	0,1148	76	Bwd Avg Bulk Rate	0,0003
35	Fwd IAT Min	0,1443	77	ECE Flag Count	0,0001
36	Fwd Header Length	0,1179	78	Bwd Avg Packets/Bulk	0,0001
37	Destination IP	0,1061	79	SYN Flag Count	0
38	Bwd Packets/s	0,0875	80	Fwd Avg Bytes/Bulk	0
39	Inbound	0,0852	81	FIN Flag Count	0
40	Init-Win-bytes-backward	0,0841	82	Bwd URG Flags	0
41	Total Backward Packets	0,0837	83	Bwd PSH Flags	0
42	Source IP	0,0835			

Table 4. Training and Testing data

Subsets	**All Attacks**		**Three Classes Classification**			**Binary Classification**		
Attacks	Training	Testing	Classes	Training	Testing	Classes	Training	Testing
Benign	4.462	1.115	Benign	4.462	1.115	Benign	4.462	1.115
Syn	258.037	64.509	Exploitation Attacks	427.467	106.867	Attack	2.811.709	702.982
UDP	154.695	38.674						
UDP-lag	14.664	3.666						
UDP-Lag	71	18						
TFTP	803.346	200.837	Reflection Attacks	2.384.242	596.061			
MSSQL	231.543	57.886						
DrDoS-SNMP	206.392	51.598						
DrDoS-DNS	202.841	50.710						
DrDoS-MSSQL	180.907	45.227						

DrDoS-NetBIOS	163.745	40.936						
NetBIOS	146.397	36.599						
DrDoS-UDP	125.391	31.348						
DrDoS-SSDP	104.433	26.108						
DrDoS-LDAP	87.184	21.796						
LDAP	76.490	12.021						
DrDoS-NTP	48.085	12.021						
Portmap	7.472	1.868						
WebDDoS	16	4						
Total All Attacks	2.816.171	704.043	Total Three Classes	2.816.171	704.043	Total Binary Analysis	2.816.171	704.043

Group 2, eleven (11) most relevant features: For the binary classification, the global results are tabulated in Table 9. Other performances indicators are tabulated in Table 5 and in Figures 13 and 14. RF outperforms the other classifiers; ET comes at the second rank and DT at the third. For the Three Classes analysis, performance results are given in Tables 9, 5 and in Figure 14. Using this subset, RF surpasses DT, ET and XGB. Using All Attacks Classification, the performance results are tabulated in Tables 9 and 5. Figures 14 and 13 also show execution times and the numbers of incorrectly classified for this subset. Using this subset, DT surpasses RF, ET and XGB. Group 3, fourteen (14) most relevant features: Performance results for this group are tabulated in Tables 9, 5, and visualized in Figures 13 and 14. When we use the Binary Classification and this subset, RF comes at first, then ET, DT and XGB. Using the Three Classes Classification and this subset, RF has the best performances, then comes DT, ET and XGB. Using the All Attacks Classification and this subset, DT and RF give the best results, then comes ET and XGB. Group 4, twenty (20) most relevant features: The performances of the classifiers are tabulated in Tables 6,7, 8, 9, and visualized in Figure 14. Using Binary Classification and this subset, RF has the highest performances, then ET, DT and XGB. Using the Three Classes classification and this subset, DT has the highest performances, then RF, ET and XGB. When we use the All Attacks Classification, DT has the highest detection rates. It predicts all attacks with precision of more than 99.95% except for Benign attacks with a precision of 97% and for UDPlag with a precision of 89%. RF also has high detection rates but it doesn't recognize WedDDoS attacks at all. ET and XGB have some lack of precision in the detection of WebDDoS, UDPlag, Benign and Portmap.

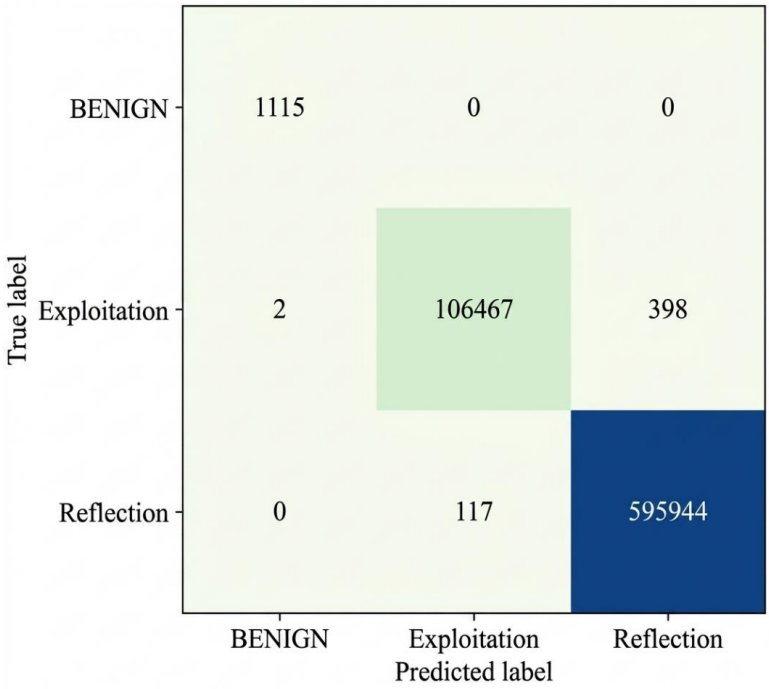


Fig. 7(a) Confusion Matrices obtained with ET. Binary Classification

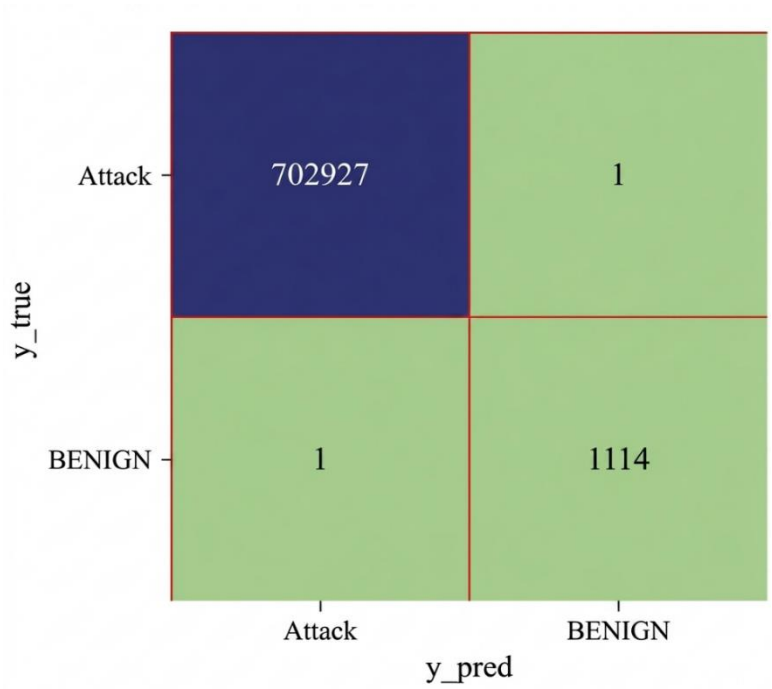


Fig. 7(b) Confusion Matrices obtained with ET Model for Three types of attacks Classification

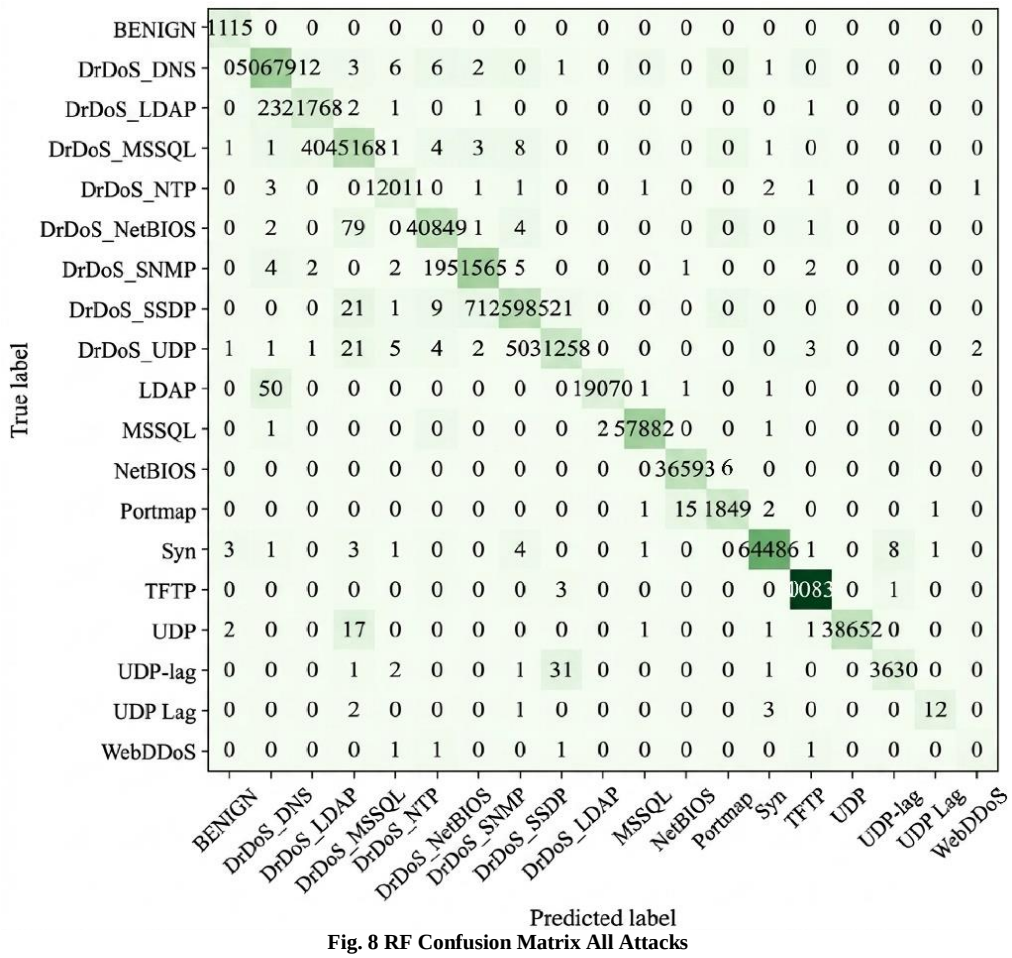


Fig. 8 RF Confusion Matrix All Attacks

Group 5, twenty-eight (28) most relevant features: Using this group, the performances of the classifiers are tabulated in Tables 8 and 9. These performances are also visualized in Figure 14. For Binary Classification and this subset, the classification performances of the four (4) classifier algorithms are close to total perfection. Random Forest has the highest detection rates with just 18 incorrectly classified attacks among 704,043 attacks, which is negligible. ET has also high detection rates with only 23 incorrectly classified attacks among 704,043 attacks. XGB and DT come at the third and fourth place with 30 and 35 incorrectly classified attacks respectively. Using the Three Classes Classification and this subset, RF has the highest detection rates with only 46 incorrectly classified instances among 704,043 attacks. DT has 49 incorrectly classified instances among 704,043 attacks. ET has 305 incorrectly classified and XGB has 417 incorrectly classified instances. Using All Attacks Classification, DT detects all attacks with precisions more than 99.5% except for Benign traffic with a precision of 97% and WebDDoS with a precision of 50%. RF detects all attacks with precisions of more than 99.5% except for Benign traffic with a precision of 97% and WebDDoS with a precision of 33%. ET and XGB have some lack in their detection rates when compared to DT and RF.

Group 6, All features (83): Performance results for Group 6 and the subset Binary Analysis Classification are given in Figures 5(a) , 7(a) , 10(a) , and 11(a) for DT, ET, RF, and XGB respectively. Figure 14 shows comparative numbers of incorrectly classified attacks for the four classifier algorithms. The performances of all classifier algorithms using all features are very good particularly for XGB which attains an accuracy, precision, recall and F1-score of 100%. As shown in Table 6, the number of incorrectly classified has decreased from 203 to 1 between Group 1 and Group 6 using DT. It has also decreased from 196 to 4 between Group 1 and Group 6 using RF and from 199 to 1 using ET and to zero (0) using XGB.

When we consider the Three Classes (labels) Analysis including Normal traffic (Benign attacks), Reflection Attacks and Exploitation Attacks, the performances are shown in confusion matrices obtained after these experiences and given in Figures 5(b), 7(b), 10(b), and 11(b) for DT, ET, RF and XGB respectively.

The global performance indicators are given in table 7. DT, RF and XGB classifiers perform better than ET classifier using this subset (Gr6) with accuracies greater than 99.99%.

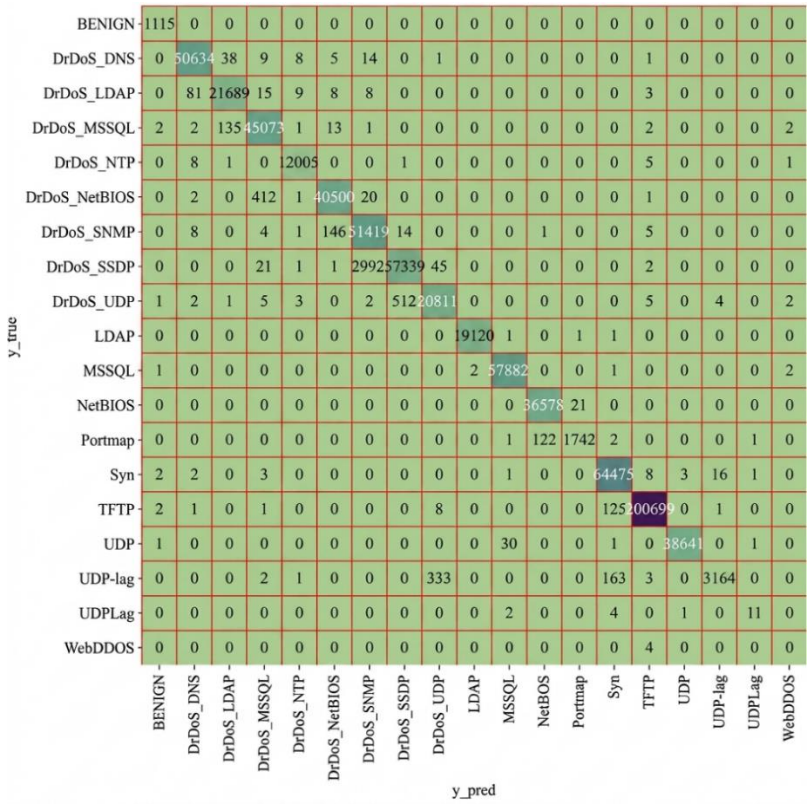


Fig. 9 ET All Attacks Confusion Matrix

But these algorithms still produce some confusion between reflection attacks and reflection attacks. Tables 7 and 8 show that just 11 reflection attacks and just 10 exploitation attacks are wrongly predicted by DT, 73 reflection attacks are predicted to be exploitation attacks and 13 exploitation attacks are predicted as reflection attacks by RF while 335 reflection

attacks are predicted to be exploitation attacks and 59 exploitation attacks are predicted as reflection attacks by XGB. ET confuses 393 reflection attacks and 117 exploitation attacks. Finally, the performances of algorithms using all features (83) are visualized in Figures 6, 9, 8 and 12 for classifiers DT, ET, RF and XGB respectively. The numbers of incorrectly classified are illustrated in Figure 14. The effects of feature selection with Mutual Information and comparison between Groups 1, 4 and 6 in terms of attacks detection errors are tabulated in Table 8.

Table 9 also shows global detection performance indicators comparison for all groups and all classifiers. RF, ET and XGB detect perfectly Normal traffic and DT predicts wrongly only 12 Benign attacks over 1,115. DT detects perfectly DrDoS-DNS attacks with only 4 incorrectly detected. RF, XGB and ET come after with respectively with 31, 50 and 76 incorrectly classified over 50,710 instances.

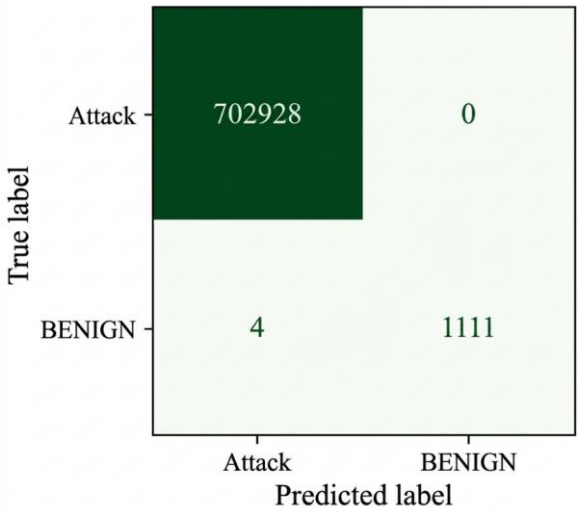


Fig. 10(a) Confusion Matrix Obtained with RF for Binary Classification

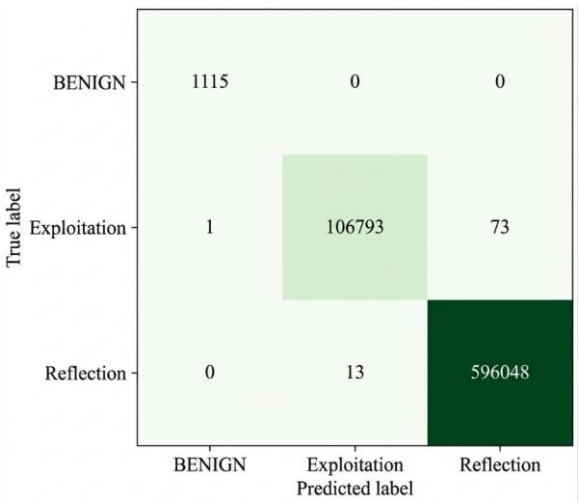


Fig. 10(b) Confusion Matrix Obtained with RF Three attacks Classification

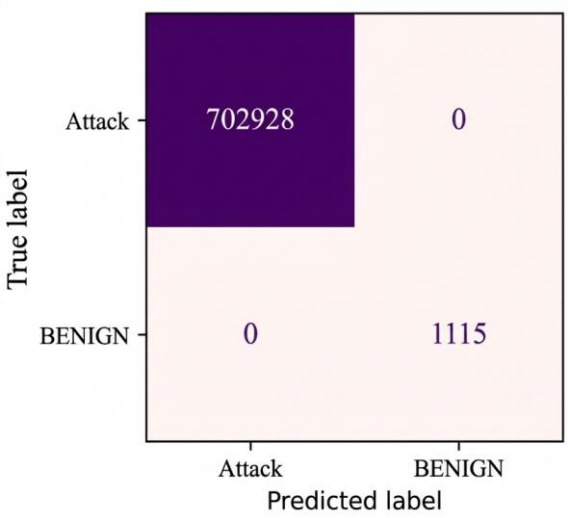


Fig. 11(a) Confusion Matrix Obtained with XGBoost Model for Binary Classification

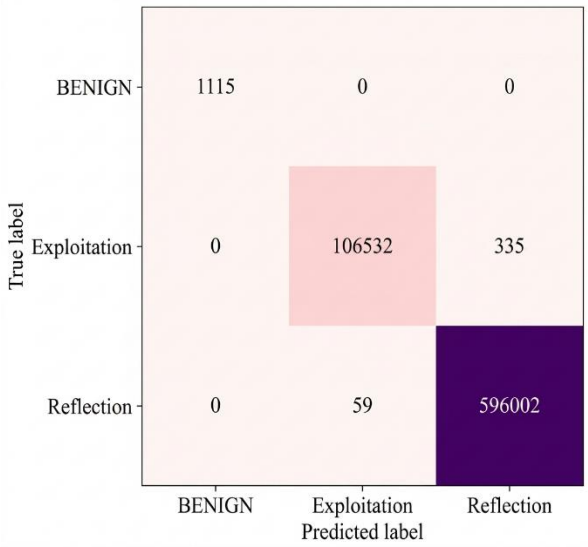


Fig. 11(b) Confusion Matrix Obtained with XGBoost Model for Three types of attacks Classification

To summarize, DT and RF surpass other classifiers for detecting DrDoS-DNS, DRDoSLDAP, DrDoS-MSSQL, DrDoS-NTP, DrDoSNetBIOS, DrDoS-SNMP, DrDoS-SSDP, DrDoSUDP, LDAP, MSSQL, NetBIOS, Syn, TFTP, UDP and UDP-Lag with precisions of more than 99.5%.

DT also predicts UDP-Lag, UDPlag with precisions of more than 99.5% and predicts WebDDoS attacks with a precision of 80%. RF is unable to predict UDPLag and WebDDoS attacks with good scores, but this may be due to the small numbers of instances used to train and test the classifier.

7.3 Comparison against some Baseline Methods

We present in Table 11, a comparison of our global performance results against some baseline previous studies.

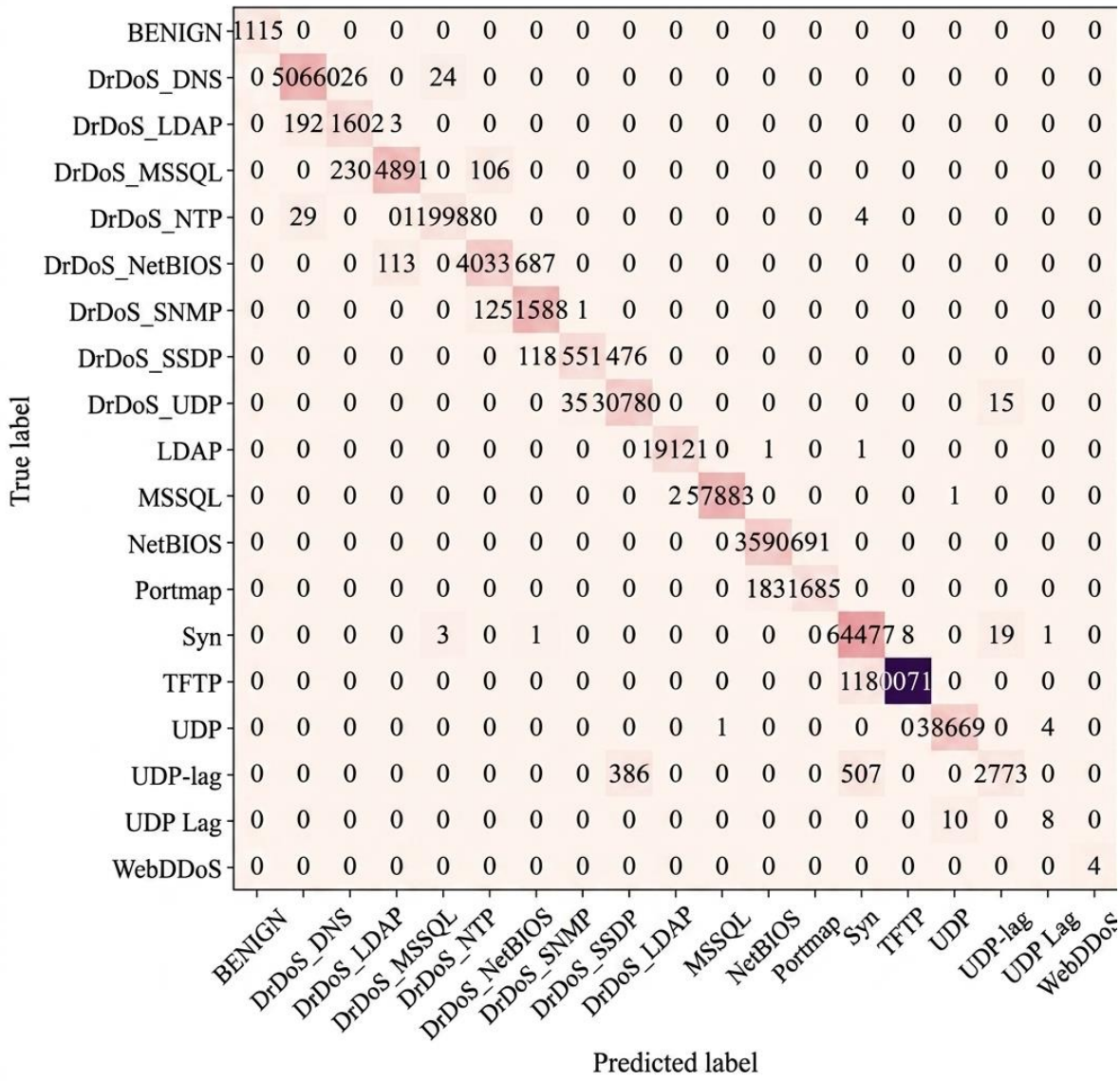


Fig. 12 XGBoost Confusion Matrix All Attacks

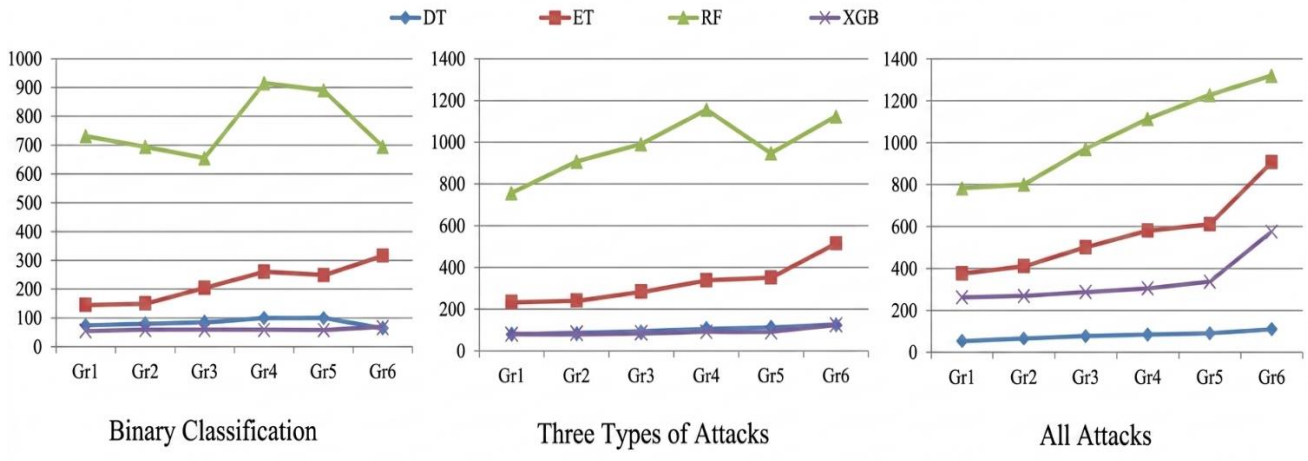


Fig. 13 Execution times in seconds

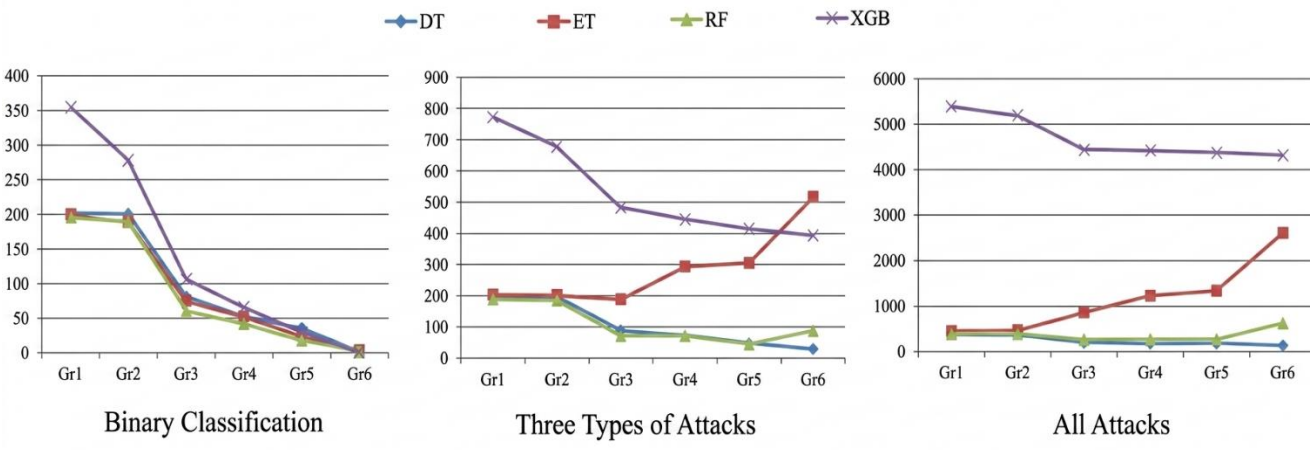


Fig. 14 Comparison of incorrectly classified

Table 5. Execution Times by Classifier in seconds

Classification	Classifiers	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
Binary Classification	Decision Tree	71	76	83	99	99	64
	Extra Trees	144	150	203	259	250	316
	Random Forest	730	693	657	914	891	695
	XGBoost	52	56	56	56	58	67
Three Attacks Classification	Decision Tree	78	78	90	106	109	124
	Extra Trees	232	235	282	338	349	511
	Random Forest	759	752	846	966	947	958
	XGBoost	78	84	83	89	90	127
All Attacks Classification	Decision Tree	58	70	79	87	94	111
	Extra Trees	371	411	504	580	612	903
	Random Forest	784	803	969	112	1231	1318
	XGBoost	263	270	285	306	337	576

Table 6: Global performance comparison between Groups 1 and 6

	Decision Tree		Extra Trees		Random Forest		XGBoost	
Groups	Gr 1	Gr 6	Gr 1	Gr 6	Gr 1	Gr 6	Gr 1	Gr 6
Selected Features	9	83	9	83	9	83	9	83
True Positives	702.834	702.928	702.836	702.928	702.835	702.928	702.782	702.928
True Negatives	1.006	1.115	1.008	1.115	1.012	1.111	906	1.115
False Positives	94	0	92	0	93	0	146	0
False Negatives	192	1	107	0	103	4	209	0
Accuracy %	96	100	100	100	100	100	100	100
Precision %	95	100	96	100	96	100	93	100
Recall %	95	100	95	100	95	100	91	100
F1-Score %	95	100	96	100	96	100	92	100

Table 7. Comparative results between Groups 1, 4 and 6 for Three Classes Analysis

	Decision Tree			Extra Trees			Random Forest			XGB		
Groups	Gr 1	Gr 4	Gr 6	Gr 1	Gr 4	Gr 6	Gr 1	Gr 4	Gr 6	Gr 1	Gr 4	Gr 6
Selected Features	9	20	83	9	20	83	9	20	83	9	20	83
TP Benign	1.036	1.084	1.109	1.037	1.087	1.115	1.043	1.095	1.115	910	1.087	1.115
TP	106.826	106.834	106.855	106.822	106.696	106.467	106.828	106.836	106.793	106.503	106.529	106.532

Exploitation												
TP Reflection	595.982	596032	596050	595984	595967	595944	595948	595039	596048	595857	505982	596002
FP Exploit	16	6	0	10	7	0	10	4	0	143	13	0
FP Reflection	63	25	0	68	21	0	62	16	0	62	15	0
FP Refl-Expl	22	10	11	27	165	398	22	26	73	339	331	335
FN Exploit	19	3	1	18	6	2	17	5	1	25	7	0
FN Reflect.	77	17	1	70	14	0	72	8	0	129	21	0
FN Expl-Refl	2	12	10	7	80	117	5	14	13	75	58	59
Preci.Expl.%	100	100	100	100	100	100	100	100	100	100	100	100
Preci. Beng%	92	99	100	92	98	92	92	99	100	86	97	100
Preci,Refl.%	100	100	100	100	100	100	100	100	100	100	100	100
Average Precision %	97	99	100	97	99	97	97	100	100	100	100	100
Accuracy %	100	100	100	100	100	100	100	100	100	100	100	100
Recall %	98	99	99	98	97	98	98	98	100	82	97	100
F1-Score %	97	99	100	98	98	98	98	99	100	84	97	100

Table 8. Comparative numbers of incorrectly classified between Groups 1, 4 and 6

	Decision Tree			**Extra Trees**			**Random Forest**			**XGBoost**		
Attacks	**Gr 1**	**Gr 4**	**Gr 6**	**Gr 1**	**Gr 4**	**Gr 6**	**Gr 1**	**Gr 4**	**Gr 6**	**Gr 1**	**Gr 4**	**Gr 6**
Benign	90	37	12	75	20	0	79	12	0	193	22	0
DrDoS-DNS	7	5	4	14	42	76	11	17	31	100	45	50
DrDoS-LDAP	4	10	11	4	27	107	4	8	18	574	204	194
DrDoS-MSSQL	15	6	6	39	80	154	22	29	59	504	347	336
DrDoS-NTP	22	11	2	24	14	17	24	10	10	48	43	33
DrDoS-NetBIOS	22	10	10	45	252	421	23	22	87	698	603	600
DrDoS-SNMP	13	21	14	28	66	84	15	25	33	5	9	13
DrDoS-SSDP	28	21	24	53	132	364	29	40	123	1193	743	594
DrDoS-UDP	25	24	24	37	170	530	30	39	90	38	485	568
LDAP	1	2	2	2	3	3	2	3	53	2	1	2
MSSQL	2	1	1	4	5	5	4	4	4	11	5	3
NetBIOS	6	6	6	6	15	23	6	6	6	813	742	693
Portmap	21	3	22	6	43	124	4	10	19	10	95	183
Syn	71	15	7	74	33	34	74	18	27	20	16	32
TFTP	38	8	1	34	81	138	36	7	10	221	132	127
UDP	2	1	2	4	15	33	3	2	22	14	10	5
UDP-lag	5	7	6	10	226	502	6	12	36	935	912	893
UDPlag	1	1	1	3	4	7	3	4	6	9	11	11
WebDDoS	1	1	0	2	3	4	2	4	4	3	1	0
Totals	374	190	155	464	1231	2626	377	272	638	5391	4426	4337

Table 9. Comparative results for Feature Selection with mutual information experiments

Algorithms	**Classification**	**Performance metrics**	**Gr 1**	**Gr 2**	**Gr 3**	**Gr 4**	**Gr 5**	**Gr 6**
Decision Tree	Binary Analysis	Execution Time	71	76	83	99	99	64
		Accuracy %	100	100	100	100	100	100
		F1-Score %	95	95	95	95	95	95
	3 Classes Analysis	Execution Time	78	78	90	106	109	124
		Accuracy %	100	100	100	100	100	100
		F1-Score %	97	97	99	99	100	100

	All Attacks Analysis	Execution Time	58	70	79	87	94	11
		Accuracy %	100	100	100	100	100	100
		F1-Score %	97	97	97	97	97	97
Extra Trees	Binary Analysis	Execution Time	144	150	203	259	250	316
		Accuracy %	100	100	100	100	100	100
		F1-Score %	96	96	96	96	96	96
	3 Classes Analysis	Execution Time	232	235	282	338	349	511
		Accuracy %	100	100	100	100	100	100
		F1-Score %	98	98	99	99	100	100
	All Attacks Analysis	Execution Time	371	411	504	580	612	903
		Accuracy %	100	100	100	100	100	100
		F1-Score %	95	95	95	95	95	95
Random Forest	Binary Analysis	Execution Time	730	693	657	914	891	695
		Accuracy %	100	100	100	100	100	100
		F1-Score %	96	96	99	98	100	100
	3 Classes Analysis	Execution Time	752	759	846	966	947	958
		Accuracy %	100	100	100	100	100	100
		F1-Score %	98	98	99	99	100	100
	All Attacks Analysis	Execution Time	784	803	969	1112	1231	1318
		Accuracy %	100	100	100	100	100	100
		F1-Score %	96	96	97	94	95	93
XGBoost	Binary Analysis	Execution Time	52	56	56	56	58	67
		Accuracy %	100	100	100	100	100	100
		F1-Score %	92	94	98	99	99	100
	3 Classes Analysis	Execution Time	78	84	83	89	90	127
		Accuracy %	100	100	100	100	100	100
		F1-Score %	94	96	99	99	100	100
	All Attacks Analysis	Execution Time	263	270	285	306	337	576
		Accuracy %	99	99	99	99	99	99
		F1-Score %	91	92	90	94	94	95

Table 10. Ranking of classifier algorithms

Groups	Subsets	Rank 1	Rank 2	Rank 3	Rank 4
Group 1	Binary	DT	ET	RF	XGB
	3 Class	RF	DT	ET	XGB
	All	DT	RF	ET	XGB
Group 2	Binary	RF	ET	DT	XGB
	3 Class	RF	ET	DT	XGB
	All	DT	RF	ET	XGB
Group 3	Binary	RF	ET	DT	XGB
	3 Class	RF	DT	ET	XGB
	All	DT	RF	ET	XGB
Group 4	Binary	RF	ET	DT	XGB
	3 Class	DT	RF	ET	XGB
	All	DT	RF	ET	XGB
Group 5	Binary	RF	ET	XGB	DT
	3 Class	RF	DT	ET	XGB
	All	DT	RF	ET	XGB
Group 6	Binary	XGB	DT	ET	RF
	3 Class	DT	RF	XGB	ET
	All	DT	RF	ET	XGB

Table 11. Global results comparison against some base line studies

Year	Authors	Datasets	Algorithms and methods	Accuracy
2024	Tamara Zhukabayeva et al. [29]	N-BaIoT	XGB , RF	99.21%
2024	S.A. Mohamed et al. [33]	CICDDoS2019	SGB,KNN, RF, SVM and NB	53% to 99%
2024	A.A. Alashhab et al. [32]	CICDDoS2019, InSDN, Custom Dataset	NB, Passive-Aggressive, MLP Ensemble Model	98.2% to 99.26%
2024	Yongqiang Shang [34]	NSL-KDD	NB, RF , PCA , LVQ	
2023	H.B.D.M. Sharif et al. [39]	CICIDS2017 , CICDDoS2019	DT , MRMR, GA	99.9%
2023	M.A. Hossain and M. Saiful Islam [80]	CCC, CICIDS,ICX , CU13,Bot-IoT	PCA , MI , SMOTE	99%
2022	M.Alduailij et al. [5]	CICIDS2017 , CICDDoS2019	MI , RFFI , KNN	99.9%
2022	Sambangi and Aljawarneh [41]	IoT DoS	SWASTHIKA	
2021	E.O. Omuya et al. [42]	Breast Cancer dataset	PCA-IG , NB	20 selected features
2021	Dhindsa et al. [43]	Private dataset (32779 microbe images)	Otsu, ISO Data, MI, KNN MLP, QDA, LR and SVM	88% , 81% and 98.1%
2020	Kurniabudi et al. [70]	CICIDS2017	RF, BN , RT , J48	0.33 to 0.99
2025	Proposed	CICDDoS2019	Mutual Information, DT, ET , RF, XGB	100 %

8. Conclusions

In this work, after a deep analysis of the CICDDoS2019 dataset, we have demonstrated and proved by experimentation the effects of using machine learning techniques coupled with Mutual Information feature selection to improve DDoS attacks intrusion detection performances. We have opted for Mutual Information because it's able to detect any kind of relationship between features, computationally efficient, and resistant to over fitting. We have experimented with four (4) machine learning classier algorithms over six groups of features and three (classification subsets).

We have ranked the classifier algorithms according to their attacks detection rates and used the ranking to create six groups of features according to their cumulated Mutual Information. When we use a Binary Classification subset, RF outperforms using Group 2, Group 3, Group 4 and Group 5. DT outperforms With Group 1 and XGB with Group 6. Using the Three Classes subset, RF outperforms using Group1, Group 2, Group 3 and Group 5. DT outperforms with Groups 4 and 6. Using the All-Attacks subset, DT and RF outperform other classifiers for all groups. There is an illustration of the effects of feature selection with Mutual Information on execution times and on intrusion detection performances. The final results show that the optimum is reached with group 5, twenty-eight (28) features. In fact, with this group, we obtain fast execution times and low numbers of incorrectly classified.

Finally, as future works we are planning to do further research work on feature selection by the use of Deep Learning methods such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) Auto-Encoders, or Large Language Models.

Conflicts of Interest

All authors declare that they have no conflicts of interest regarding the publication of this paper.

Funding Statement

Al authors state that this article is not funded and not supported by any association.

References

[1] "*Cisco 2018 Annual Cybersecurity Report*," Technical Report, Technical Report by Cisco systems, 2018. [Publisher Link]

[2] Statista - The Statistics Portal, Statista, 2020. [Online]. Available: www.statista.com.

[3] Muhammad Ashfaq Khan, "HCRNNIDS: Hybrid Convolutional Recurrent Neural Network-Based Network Intrusion Detection System," *Processes*, vol. 9, no. 5, pp. 1-14, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Yuanyuan Wei et al., "AE-MLP: A Hybrid Deep Learning Approach for DDoS Detection and Classification," *IEEE Access*, vol. 9, pp. 146810-146821, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Mona Alduailij et al., "Machine-Learning-Based DDoS Attack Detection Using Mutual Information and Random Forest Feature Importance Method," *Symmetry*, vol. 14, no. 6, pp. 1-15, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Swathi Sambangi, and Lakshmeeswari Gondi, "A Machine Learning Approach for DDoS (Distributed Denial of Service) Attack Detection Using Multiple Linear Regression," *Proceedings*, vol. 63, no. 1, pp. 1-12, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[7] Tasnuva Mahjabin et al., "A Survey of Distributed Denial-of-Service Attack, Prevention, and Mitigation Techniques," *International Journal of Distributed Sensor Networks*, vol. 13, no. 12, pp. 1-33, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[8] Rutika S. Chaudhari, and Girish Talmale, "A Review on Detection Approaches for Distributed Denial of Service Attacks," *2019 International Conference on Intelligent Sustainable Systems (ICISS)*, Palladam, India, pp. 323-327, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[9] Muhammad Naveed et al., "A Deep Learning-Based Framework for Feature Extraction and Classification of Intrusion Detection in Networks," *Wireless Communications and Mobile Computing*, vol. 2022, no. 1, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ziadoon Kamil Maseer et al., "Benchmarking of Machine Learning for Anomaly Based Intrusion Detection Systems in the CICIDS2017 Dataset," *IEEE Access*, vol. 9, pp. 22351-22370, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Lirim Ashiku, and Cihan Dagli, "Network Intrusion Detection System Using Deep Learning," *Procedia Computer Science*, vol. 185, pp. 239-247, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Zahra Jadidi et al., "Flow-Based Anomaly Detection Using Neural Network Optimized with GSA Algorithm," *2013 IEEE 33rd International Conference on Distributed Computing Systems Workshops*, Philadelphia, PA, USA, pp. 76-81, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[13] Ansam Khraisat et al., "Survey of Intrusion Detection Systems: Techniques, Datasets and Challenges," *Cybersecurity*, vol. 2, no. 1, pp. 1-22, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[14] Bo Sun et al., "Intrusion Detection Techniques in Mobile Ad Hoc and Wireless Sensor Networks," *IEEE Wireless Communications*, vol. 14, no. 5, pp. 56-63, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[15] Youssef Regragui et al., "Impact Evaluation of Feature Selection Algorithms on Machine Learning-Based Intrusion Detection," *2024 11th International Conference on Wireless Networks and Mobile Communications (WINCOM)*, Leeds, United Kingdom, pp. 1-6, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[16] Xuan-Ha Nguyen, and Kim-Hung Le, "Robust Detection of Unknown DoS/DDoS Attacks in IoT Networks Using a Hybrid Learning Model," *Internet of Things*, vol. 23, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Ansam Khraisat, and Ammar Alazab, "A Critical Review of Intrusion Detection Systems in the Internet of Things: Techniques, Deployment Strategy, Validation Strategy, Attacks, Public Datasets and Challenges," *Cybersecurity*, vol. 4, no. 1, pp. 1-27, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[18] Yi Xie, and Shun-Zheng Yu, "A Large-Scale Hidden Semi-Markov Model for Anomaly Detection on User Browsing Behaviors," *IEEE/ACM Transactions on Networking*, vol. 17, no. 1, pp. 54-65, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[19] Ognjen Joldzic, Zoran Djuric, and Pavle Vuletic, "A Transparent and Scalable Anomaly-Based DoS Detection Method," *Computer Networks*, vol. 104, pp. 27-42, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[20] Maulik Gohil, and Sathish Kumar, "Evaluation of Classification Algorithms for Distributed Denial of Service Attack Detection," *2020 IEEE Third International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, Laguna Hills, CA, USA, pp. 138-141, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[21] Julius Jow, Yang Xiao, and Wenlin Han, "A Survey of Intrusion Detection Systems in Smart Grid," *International Journal of Sensor Networks*, vol. 23, no. 3, pp. 170-186, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[22] Elijah M. Maseno, Zenghui Wang, and Hongyan Xing, "A Systematic Review on Hybrid Intrusion Detection System," *Security and Communication Networks*, vol. 2022, no. 1, pp. 1-23, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Sulaiman Alhaidari, and Mohamed Zohdy, "Hybrid Learning Approach of Combining Cluster-Based Partitioning and Hidden Markov Model for IoT Intrusion Detection," *Proceedings of the 2019 3rd International Conference on Information System and Data Mining*, pp. 27-31, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[24] B. Geluvaraj, P. M. Satwik, and T. A. Ashok Kumar et al., "The Future of Cybersecurity: Major Role of Artificial Intelligence, Machine Learning, and Deep Learning in Cyberspace," *International Conference on Computer Networks and Communication Technologies: ICCNCT 2018*, Singapore, pp. 739-747, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[25] Bilgehan Arslan, Sedef Gunduz, and Seref Sagiroglu, "A Review on Mobile Threats and Machine Learning Based Detection Approaches," *2016 4th International Symposium on Digital Forensic and Security (ISDFS)*, Little Rock, AR, USA, pp. 7-13, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[26] Kamran Shaukat et al., "A Survey on Machine Learning Techniques for Cyber Security in the Last Decade," *IEEE Access*, vol. 8, pp. 222310-222354, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[27] Mouhammd Al-Kasassbeh et al., "Feature Selection Using a Machine Learning to Classify a Malware," *Handbook of Computer Networks and Cyber Security: Principles and Paradigms*, pp. 889-904, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[28] Kahraman Kostas, "Anomaly Detection in Networks Using Machine Learning," *Research Proposal*, vol. 23, pp. 1-70, 2018. [Google Scholar]

[29] Tamara Zhukabayeva et al., "Enhancing IoT Security: Effective Botnet Attack Detection through Machine Learning," *Procedia Computer Science*, vol. 241, pp. 421-426, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[30] Abdussalam Ahmed Alashhab et al., "Enhancing DDoS Attack Detection and Mitigation in SDN Using an Ensemble Online Machine Learning Model," *IEEE Access*, vol. 12, pp. 51630-51649, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[31] Akindele S. Afolabi, and Olubunmi A. Akinola, "Network Intrusion Detection Using Knapsack Optimization, Mutual Information Gain, and Machine Learning," *Journal of Electrical and Computer Engineering*, vol. 2024, no. 1, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[32] Abdussalam Ahmed Alashhab et al., "Ensemble Based Detection Model for DDoS Attacks in SDNs Using Advanced Feature Selection," *2024 17th International Conference on Signal Processing and Communication System (ICSPCS)*, Surfers Paradise, Australia, pp. 1-5, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[33] Ahmed Mohamed Salama, Mohamed AbdElAzim Mohamed, and Eman AbdElhalim, "Enhancing Network Security in IoT Applications through DDoS Attack Detection Using ML," *Mansoura Engineering Journal*, vol. 49, no. 3, pp. 1-21, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[34] Yongqiang Shang, "Prevention and Detection of DDoS Attack in Virtual Cloud Computing Environment Using Naive Bayes Algorithm of Machine Learning," *Measurement: Sensors*, vol. 31, pp. 1-9, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[35] Md. Alamgir Hossain, and Md. Saiful Islam, "Enhancing DDoS Attack Detection with Hybrid Feature Selection and Ensemble-Based Classifier: A Promising Solution for Robust Cybersecurity," *Measurement: Sensors*, vol. 32, pp. 1-12, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[36] Dyari Mohammed Sharif, and Hakem Beitollahi, "Detection of Application-Layer DDoS Attacks Using Machine Learning and Genetic Algorithms," *Computers & Security*, vol. 135, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[37] Siriporn Chimphlee, and Witcha Chimphlee, "Machine Learning to Improve the Performance of Anomaly-Based Network Intrusion Detection in Big Data," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 2, pp. 1106-1119, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[38] Mohammad Najafimehr, Sajjad Zarifzadeh, and Seyedakbar Mostafavi, "DDoS Attacks and Machine-Learning-Based Detection Methods: A Survey and Taxonomy," *Engineering Reports*, vol. 5, no. 12, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[39] Mohamed Riadh Kadri et al., "Survey and Classification of Dos and DDos Attack Detection and Validation Approaches for IoT Environments," *Internet of Things*, vol. 25, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[40] Erick Odhiambo Omuya, George Okeyo, and Michael Kimwele, "Sentiment Analysis on Social Media Tweets Using Dimensionality Reduction and Natural Language Processing," *Engineering Reports*, vol. 5, no. 3, pp. 1-14, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[41] Swathi Sambangi Lakshmeeswari Gondi, and Shadi Aljawarneh, "A Feature Similarity Machine Learning Model for DDoS Attack Detection in Modern Network Environments for Industry 4.0," *Computers and Electrical Engineering*, vol. 100, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[42] Erick Odhiambo Omuya, George Onyango Okeyo, and Michael Waema Kimwele, "Feature Selection for Classification Using Principal Component Analysis and Information Gain," *Expert Systems with Applications*, vol. 174, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[43] Anaahat Dhindsa et al., "An Improvised Machine Learning Model Based on Mutual Information Feature Selection Approach for Microbes Classification," *Entropy*, vol. 23, no. 2, pp. 1-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[44] Md Al-Imran, and Shamim H. Ripon, "Network Intrusion Detection: An Analytical Assessment Using Deep Learning and State-of-the-Art Machine Learning Models," *International Journal of Computational Intelligence Systems*, vol. 14, no. 1, pp. 1-20, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[45] Mohammed Al-Sarem et al., "An Aggregated Mutual Information Based Feature Selection with Machine Learning Methods for Enhancing IoT Botnet Attack Detection," *Sensors*, vol. 22, no. 1, pp. 1-20, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[46] Majid Torabi et al., "A Review on Feature Selection and Ensemble Techniques for Intrusion Detection System," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, pp. 538-553, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[47] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," *In Proceedings of the 4th International Conference on Information Systems Security and Privacy ICISSP*, Funchal, Madeira, Portugal, vol. 1, pp. 108-116, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[48] Arash Habibi Lashkari et al., "Characterization of Tor Traffic Using Time Based Features," *In Proceedings of the 3rd International Conference on Information Systems Security and Privacy ICISSP*, Porto, Portugal, vol. 1, pp. 253-262, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[49] Mehmud Abliz, "*Internet Denial of Service Attacks and Defense Mechanisms*," Technical Report, University of Pittsburgh, pp. 1-50, 2011. [Google Scholar]

[50] Alvin Huseinović et al., "A Survey of Denial-of-Service Attacks and Solutions in the Smart Grid," *IEEE Access*, vol. 8, pp. 177447-177470, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[51] Monowar H. Bhuyan, D. K. Bhattacharyya, and Jugal K. Kalita, "An Empirical Evaluation of Information Metrics for Low-Rate and High-Rate DDoS Attack Detection," *Pattern Recognition Letters*, vol. 51, pp. 1-7, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[52] Abebe Abeshu Diro, and Naveen Chilamkurti, "Distributed Attack Detection Scheme Using Deep Learning Approach for Internet of Things," *Future Generation Computer Systems*, vol. 82, pp. 761-768, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[53] Omer Yoachimik, and Jorge Pacheco, DDoS Threat Report for 2023 Q4, Cloudflare, 2023. [Online]. Available: https://blog.cloudflare.com/ddos-threat-report-2023-q4/

[54] Rocky K. C. Chang, "Defending against Flooding-Based Distributed Denial-of-Service Attacks: A Tutorial," *IEEE Communications Magazine*, vol. 40, no. 10, pp. 42-51, 2002. [CrossRef] [Google Scholar] [Publisher Link]

[55] Mohammad Masdari, and Marzie Jalali, "A Survey and Taxonomy of DoS Attacks in Cloud Computing," *Security and Communication Networks*, vol. 9, no. 16, pp. 3724-3751, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[56] Clément Boin et al., "Scale Matters: A Comparative Study of Datasets for DDoS Attack Detection in CSP Infrastructure," *2023 IEEE 12th International Conference on Cloud Networking (CloudNet)*, Hoboken, NJ, USA, pp. 27-35, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[57] Clément Boin et al., "One Year of DDoS Attacks against a Cloud Provider: An Overview," *2022 4th International Conference on Advances in Computer Technology, Information Science and Communications (CTISC)*, Suzhou, China, pp. 1-5, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[58] Kameswari Kotapati et al., "A Taxonomy of Cyber Attacks on 3G Networks," *International Conference on Intelligence and Security Informatics*, pp. 631-633, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[59] Iman Sharafaldi et al., "Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy," *2019 International Carnahan Conference on Security Technology (ICCST)*, Chennai, India, pp. 1-8, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[60] Datasets, SCVIC-TS-2022: Network intrusion data with original raw network packets, IEEEDataPort, 2023.[Online]. Available: https://ieee-dataport.org/documents/scvic-ts-2022-network-intrusion-data-original-raw-network-packets

[61] Liang Xiao et al., "IoT Security Techniques Based on Machine Learning: How Do IoT Devices Use AI to Enhance Security?," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 41-49, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[62] Inès Ben Kraiem, "*Multiple Anomaly Detection by Automatic Rule Learning in Time Series*," University of Toulouse-Jean Jaurès, pp. 1-145, 2021. [Google Scholar] [Publisher Link]

[63] Sumeet Dua, and Xian Du, *Data Mining and Machine Learning in Cybersecurity*, CRC Press, 2016. [Google Scholar] [Publisher Link]

[64] Parag Saxena, *Ultimate Machine Learning with Scikit-Learn: Unleash the Power of Scikit-Learn and Python to Build Cutting-Edge Predictive Modeling Applications and Unlock Deeper Insights Into Machine Learning*, Orange Education Pvt. Ltd., 2024. [Google Scholar] [Publisher Link]

[65] Gilles Louppe, "*Understanding Random Forests: From Theory to Practice*," PhD dissertation, Universite de Liege, 2014. [Google Scholar]

[66] Ansam Khraisat, Iqbal Gondal, and Peter Vamplew, "An Anomaly Intrusion Detection System Using C5 Decision Tree Classifier," Trends *and Applications in Knowledge Discovery and Data Mining: PAKDD 2018*, pp. 149-155, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[67] HongFang Zhou et al., "A Feature Selection Algorithm of Decision Tree Based on Feature Weight," *Expert Systems with Applications*, vol. 164, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[68] Ahmad Turmudi Zy et al., "Detecting DDoS Attacks through Decision Tree Analysis: An EDA Approach with the CIC DDoS 2019 Dataset," *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, Yogyakarta, Indonesia, pp. 202-207, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[69] Manjula C. Belavagi, and Balachandra Muniyal, "Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection," *Procedia Computer Science*, vol. 89, pp. 117-123, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[70] Kurniabudi et al., "CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection," *IEEE Access*, vol. 8, pp. 132911-132921, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[71] Rabie A. Ramadan, and Kusum Yadav, "A Novel Hybrid Intrusion Detection System (IDS) for the Detection of Internet of Things (IoT) Network Attacks," *Annals of Emerging Technologies in Computing (AETiC)*, vol. 4, no. 5, pp. 61-74, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[72] Gerard Drapper Gil et al., "Characterization of Encrypted and VPN Traffic using Time-Related Features," *In Proceedings of the 2nd International Conference on Information Systems Security and Privacy ICISSP*, Rome, Italy, vol. 1, pp. 407-414, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[73] Tala Talaei Khoei et al., "Ensemble Learning Methods for Anomaly Intrusion Detection System in Smart Grid," *2021 IEEE International Conference on Electro Information Technology (EIT)*, Mt. Pleasant, MI, USA, pp. 129-135, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[74] Raj Kumar Batchu, and Hari Seetha, "On Improving the Performance of DDoS Attack Detection System," *Microprocessors and Microsystems*, vol. 93, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[75] Junhong Li, "*Detection of DDoS Attacks Based on Dense Neural Networks, Autoencoders and Pearson Correlation Coefficient*," Faculty of Graduate Studies Online Theses, Dalhousie University Halifax, 2020. [Google Scholar] [Publisher Link]

[76] Wes McKinney, *Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython*, 3rd ed., O'REILLY, 2017. [Google Scholar] [Publisher Link]

[77] Pandas, Pandas - Python Data Analysis Library, 2026. [Online]. Available: https://pandas.pydata.org/.

[78] Mahbod Tavallaee et al., "A Detailed Analysis of the KDD Cup 99 Data Set," *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, Ottawa, ON, Canada, pp. 1-6, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[79] Matt Harrisson, *Learning Pandas Python Tools for Data Munging, Data Analysis, and Visualization*, WordPress, pp. 1-208, 2016. [Publisher Link]

[80] Md Alamgir Hossain, and Md Saiful Islam, "A Novel Hybrid Feature Selection and Ensemble-based Machine Learning Approach for Botnet Detection," *Scientific Reports*, vol. 13, no. 1, pp. 1-28, 2023. [CrossRef] [Google Scholar] [Publisher Link]