ISSN: 2641-2039

Global Journal of Engineering Sciences

DOI: 10.33552/GJES.2024.11.000768

Iris Publishers

Research Article

Sequential Future Selection and Machine Learning-based Firewall Logs Classification

Qazi Waqas Khan*

Department of Computer Engineering, Jeju National University, Jeju Special Self-Governing Province, Republic of Korea

***Corresponding author:** Qazi Waqas Khan, Department of Computer Engineering, Jeju National University, Jejusi 63243, Jeju Special Self-Governing Province, Republic of Korea.

Received Date: June 24, 2024

Published Date: July 02, 2024

Abstract

Analyzing firewall logs and controlling the network traffic according to end-user behaviour is very important. It helps to increase the network's security and reduce the network overhead so that necessary actions can be taken to control user traffic on the network. Machine Learning techniques are used to analyze the patterns and to take the necessary actions according to the input patterns. This study performed experiments on the firewall log files dataset, which is available publicly. Sequential Feature Selection (SFS) is used to select relevant features. Synthetic Minority Oversampling Technique is applied to re-sample the minority classes. Gradient Boosting, Support Vector Machine, and Adaboost classifiers are applied, and the results are compared with our proposed Ensemble Weighted Voting Classifier on the firewall log files dataset for classification of firewall log files.

Keywords: Machine learning; SMOTE; Firewall; log file analysis; SFS; SMOTE; Ensemble learning

Introduction

A huge amount of data is generated by the network user in every network due to technological advancements. Analysis of end-user behaviour is most important to maintain the security and management of the network. Firewalls can control the traffic non-legitimately [1], and, according to the network policy, they take the necessary action to allow or deny the network traffic. It can conceal the network schema so that it reveals a few information to outside network nodes. The firewall is a single network system or a group of two or more network systems combined to perform the operation of a firewall. The firewall preserves information from malicious attacks [2] and makes a system more reliable and secure. The system administrator of any organization configures the firewall according to the organization's rules and regulations [3]. A vast amount of audit entries per day are often generated by firewall logs. The traffic information collected is used to create a set of data [4]. Using this data set, some data mining algorithms are built to analyse firewall log files. The firewall log files are used for forensics of the network, and it helps to understand the network traffic patterns.

Machine Learning techniques are applied to classify firewall log files in the literature. In a paper [5], use Random Forest, Naïve Bayes, Artificial Neural Network and K Nearest Neighbor to classify firewall log files. In this study, we proposed a method for classifying firewall log files. The sequential feature selection method selects the best features from all feature sets. The given dataset is suffering from an imbalanced class problem, and to solve the imbalanced class problem, we use a SMOTE oversampling method. After this, we apply Gradient Boosting, AdaBoost and Support Vector Machine classifier to classify firewall log files. An ensemble-based Ensemble Weighted Voting classifier is proposed for prediction.

The rest of the paper is organized as follows: Section 2 discusses the existing studies for analyzing and classifying firewall log files.

Section 3 describes the detailed proposed methodology. Sections 4 and 5 present the results and conclusions.

Literature Review

The paper [6] explains an empirical and experimental technique for cyber log file analysis. It applies the Tabulated Feature Vector technique and Integrated Feedback mechanism. It suggests Topological Data Analysis for log analysis explaining how to build an anomaly predictor that is replicable for other types of data and files with different types of data attributes. The experimental results indicate that the proposed technique is accurate and productive for the prediction and evaluation of log anomalies.

The study [7] proposes a performance evaluation technique of quirk prediction for firewall policy. It applies machine learning techniques for attribute selection and different machine learning evaluators for determining performance. The results show that KNN has achieved higher accuracy than others. It investigates 93 rules in the model, 6 of which showed irregularity, and also further dives deeper into the rules based on expert opinion for security purposes. The research shows that anomalies can be predicted in firewall policies by using machine learning techniques that are very helpful to overcome security issues.

The paper [8] illustrates through analysis of log files to overcome the issues regarding attacks related to data like security problems, problems related to debugging, etc. by predicting the attacks of data through generating alert occurrence in any execution or system. Information can be exposed during amendment or in the early stages of system manufacturing through logs and be used for various objectives.

In a paper [9], the quality of internet service for users is improved. It proposed the Generalized Sequential Pattern (GSP) algorithm, primarily used for mining the patterns in a company's real data. GSP extracts multiple patterns from the data based on user behaviour regarding the usage of the internet to facilitate the user by providing them with enhanced or improved quality internet services. The results demonstrate useful and insightful patterns that can be used for enhancing the service quality of the internet for the users.

The paper [10] predicts firewall logs by scanning the network to either stop or permit the traffic of the network. They use the Bee Swarm Optimization (BSO) algorithm based on the attribute/feature selection strategy for scanning purposes. They merge optimal attributes/features in their work. The results of the BSO optimal attribute/feature merging algorithm provide accurate and efficient output as compared to all feature merging approaches.

In a paper [11], machine learning algorithms are used to detect malicious user requests in a DNS firewall experiment. This experiment is performed on 34 features and 90k records of real DNS logs. The quantitative finding depicts that their method detects benign and malicious domains with a range of 89 to 96% accuracy.

In a paper [12], the use of the Generalized Sequential Pattern Mining (GSP mining) algorithm helps to understand users' behavior on the Internet. The experiment was conducted on a dataset of Firewall log files in Thailand. The experimental findings show that the method's highest F-score is 0.90.

A paper [13] analyses internet log files to identify malicious attacks from different websites and uses ElasticSearch, Logstash, and Kibana for log file analysis. Furthermore, a real-time alert system was built that generated an alarm based on the conditions. In a study [14] used a Support Vector Machine (SVM) for the classification of Firewall log files. The experiment was performed on a dataset of firewall log files. The experimental result shows that SVM with RBF kernel classifies the firewall log files with an F-Score of 78%.

Proposed Methodology

This section discussed the details of the proposed method for firewall log files classification. Figure 1 shows the proposed methodology diagram for firewall log file classification. It shows that we first input the given network parameters into the pre-processing module. The pre-processing module performs the label encoding, Feature selection, re-sampling and data standardisation. These selected attributes pass to the machine learning model for classification. The output of each model is passed to the evaluation metrics to evaluate the performance of each model.

Dataset Detail

This study performed the experiment on the firewall log files dataset that is available on the UCI repository. It has 11 independent features and 65532 instances that contain information about the user's traffic details. The target class attribute is categorical and has four classes: deny, allow, drop, and reset. The target attribute of each category describes each specific task related to firewall action. The deny action represents that allows the internet traffic, and the deny represents that blocks the traffic. Drop action represents the silent drop of the traffic. Reset both actions sends transmission control protocol to server and client-side devices. The other features are source port, destination port, NAT source port, NAT destination port, bytes, bytes sent, bytes received, elapsed time in seconds, packet, packet sent, and packet received.

Pre-Processing

This study utilized label encoding and the standard scalar method to prepare the data for the machine learning model.

Label Encoding

Label encoding is a process of converting the string category into a numeric category. The purpose of performing label encoding is that most machine learning methods only work with numeric data. In this study, we performed the label encoding of the target attribute using the label encoding function. After the Label encoding, Allow the class to assign a 0, deny the class to assign a 1, drop the class to assign a 2, and Both assign a 3 category, respectively.

Feature Selection

Feature selection is a process of selecting the most important and relevant features from all features. Feature selection has two advantages: First, it helps to reduce the computational cost of the model; second, it helps to reduce the overfitting. Because irrelevant attributes cause overfitting and affect the model performance [15]. The sequential feature selection method selects the most relevant features from the firewall log files data set.

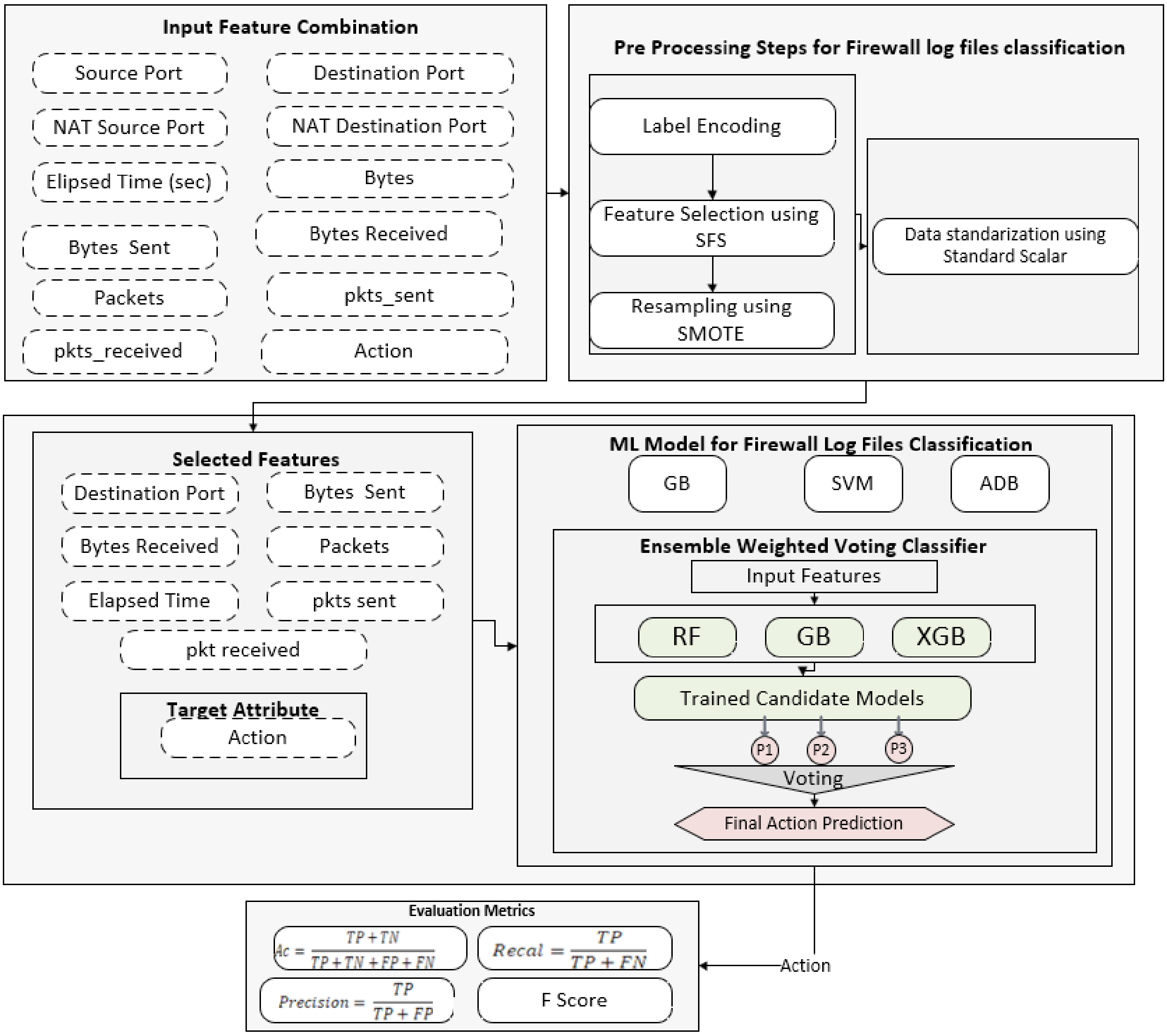

Figure 1: Proposed Methodology Diagram.

Sequential feature selection is a wrapper feature selection method that uses a greedy approach to pick the features. SFS selects the best new feature for each iteration based on the cross-validation score. It removes or adds features at a time based on the algorithm's performance until a feature subset reaches the desired number of features. This study uses the forward strategy of SFS to select the features. This method initially starts with no feature in the model, and we keep adding the features in each iteration until adding a new feature does not improve the prediction performance [16].

Synthetic Minority Oversampling Technique

Re-sampling is the process of drawing a new sample from the data from the original samples. The purpose of using the re sampling technique is to balance the data set. In an imbalanced classification problem, the class distribution is biased towards some categories, such as action attribute reset, where both categories have few instances and allow the category to have more than half of instances in a given data set. This study uses the SMOTE oversampling technique to re-sample the instances of the Action class label. SMOTE re-samples the instances by selecting the instances near each other in a feature space. It draws a line in the feature space between the data points and, after this, draws a new data point along that line [17].

Standard Scalar

The standard scalar method shown in Eq. 1 standardizes the firewall log files data set.

Z=X-mean/SD

X is an input feature, the mean is an average of features, and SD is the standard deviation of input features.

Machine Learning Method

This study applies Support Vector Machine, Adaboost, Gradient Boosting and the proposed Ensemble Weighted Voting Classifier for firewall log file classification.

Adaboost [18] is an ensemble learning classifier that combines many weak learners to build a strong prediction model. Adaboost classifier has four main steps that are performed in both classification and regression problems. A first decision stump is built on the training data, and in the second step, it creates a decision stump of different variables and observes the performance of each stump to its target classes. After this, it assigned more weight to the classifier that incorrectly classified the sample. It reiterates the steps until the maximum iteration criteria are met.

Support Vector Machine [19] is a supervised Learning technique that can be used in classification and regression problems. Support Vector Machine built a decision boundary that separates the 11-dimension feature space of the firewall log file dataset into action category classes. This decision boundary is used to correct the category of new input instances. The goal of SVM is to draw a decision boundary between data points by a large marginal hyperplane. This study uses the rbf kernel.

Gradient Boosting [20] is an ensemble learning model that is an improved version of the AdaBoost model. It also creates multiple prediction models like Adaboost to build a strong prediction model. The difference is that in GB, we do not provide the n estimator based on our choice. It used a decision tree as a default n estimator. Gradient Boosting has three main elements: weak model, additive model, and loss function.

Ensemble Learning is a machine learning strategy combining the performance of multiple classifiers. Ensemble Voting is a strategy in which we combine the performance of multiple classifiers based on voting. However, in the Ensemble Weighted Voting strategy [21], we assigned the weight to each classifier based on the performance. In ensemble voting, each classifier contributes equally. The purpose of using a weighted voting strategy is to assign the weight based on the classifier's performance. In this study, we use Random, Forest, Gradient Boosting, and Extreme Gradient Boosting as a based learner. The purpose of using Random Forest, Gradient Boosting and Xtreme Gradient Boosting is that they performed better in the case of Tabular data.

Evaluation Metrics

F score, accuracy, recall, and precision metrics are used to evaluate the performance of each method.

Results and Discussion

Gradient Boosting, Support Vector Machine. The Adaboost and Ensemble Weighted Voting Classifier is applied on firewall log dataset for classification of firewall log files. The prediction performance of Ensemble Weighted Voting Classifier is high as compared to other methods. Table 1 shows the results of prediction method for classification of firewall log files. The prediction result of this method is better as compared to the other existing method.

Table 1 shows that Ensemble Weighted Voting Classifier has the highest results in precision, accuracy, recall, and F score, which demonstrate that this method correctly classifies the firewall log files into their respective actions. Adaptive Gradient Boosting correctly classifies the log files compared to Gradient Boosting and Support Vector Machine.

Table 1: Result comparison of machine learning method for classification of firewall log files.

Method	Accuracy	Precision	Recall	F score
Adaboost	95.53	96.85	95.52	95.99
Gradient Boosting	92.98	92.23	92.9	92.6
Support Vector Machine	93.5	92.711	93.51	92.92
Ensemble Weighted Voting Classifier	99.72	99.7	99.73	99.718

Figure 3 shows the comparison graph of machine learning classifiers based on accuracy and F score. It shows that the EWV classifier has the highest prediction accuracy and F score compared to other methods. Adaboost has the second highest prediction accuracy and F score. It concludes that the Ensemble Weighted Voting classifier classifies all instances in their respective classes.

Figure 3 shows the comparison graph of machine learning classifiers based on Precision and Recall. It shows that the EWV classifier has the highest Precision and Recall score compared to other methods. This demonstrates that the EWV classifier has a low miss classification error compared to other methods.

Citation: Qazi Waqas Khan*. Sequential Future Selection and Machine Learning-based Firewall Logs Classification. Glob J Eng Sci. 11(4): 2024. GJES.MS.ID.000768. **DOI:** 10.33552/GJES.2024.11.000768.

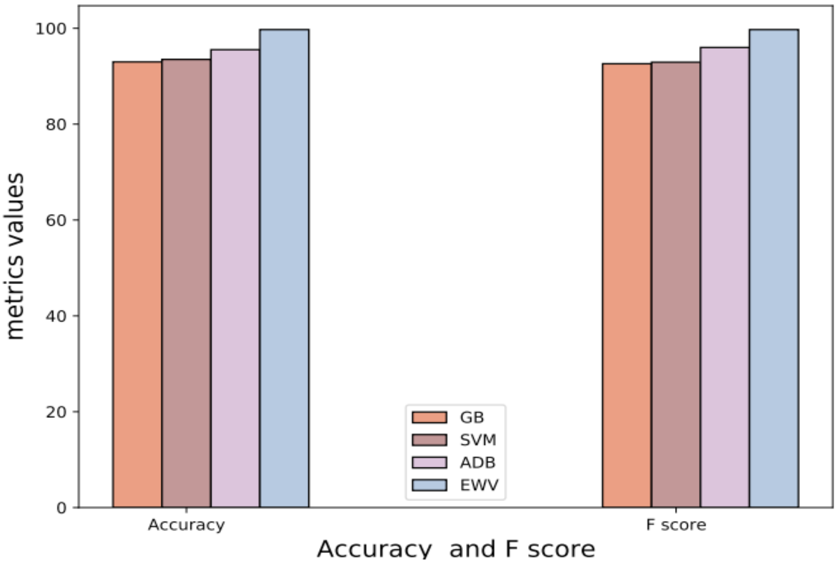

Figure 2: Comparative Analysis of Machine Learning Models based on Accuracy and Fscore.

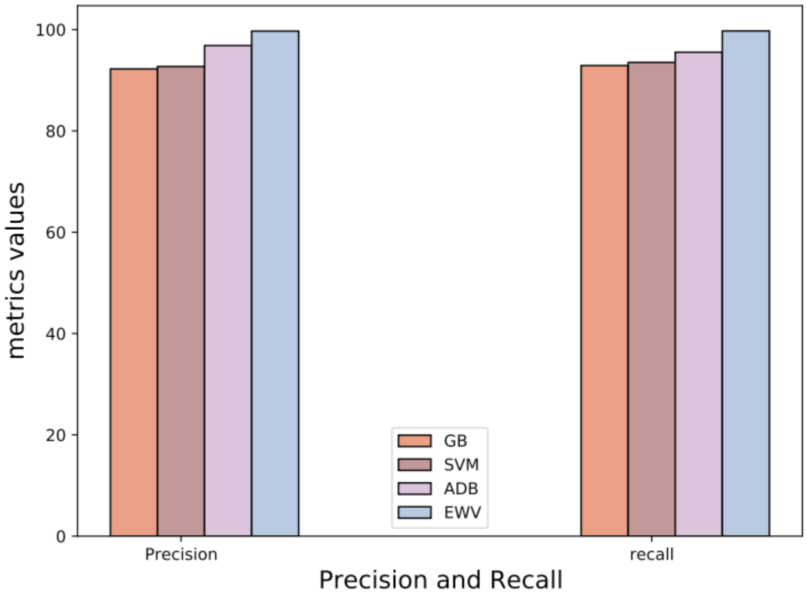

Figure 3: Comparative Analysis of Machine Learning Models based on Precision and Recall.

Conclusion

In the modern world, everything is connected to everything, and the security of the devices is mandatory. The firewall log file analysis is essential; it helps to monitor and control the internet traffic flow. It helps to understand traffic behaviour and patterns. This study experimented on a publicly available data set of firewalls on the UCI Machine Learning repository. This study first selects the relevant features using SFS feature selection and re-samples the data samples using the SMOTE re-sampling method. Gradient Boosting, Support Vector Machine, Adaboost and Ensemble weighted Voting classifier is applied. This method will help to take the necessary action on the firewall to control traffic flow better. In the future, other feature selection methods, such as the wrapper and optimization methods, may apply to select the more relevant features. The bagging ensemble method can be integrated with the deep tabular method for firewall log file classification.

Acknowledgement

None.

Conflict of Interest

The author declares no conflict of interest.

References

1. Kizza, Joseph Migga (2024) Firewalls. Guide to Computer Network Security. Cham: Springer International Publishing 265-294.
2. Dawadi, Babu R, Bibek Adhikari, Devesh Kumar Srivastava (2023) Deep learning technique-enabled web application firewall for the detection of web attacks. Sensors 23(4): 2073.
3. Md Shamimul Islam, Nayan Kumar Datta, Md Imran Hossain Showrov, Md Mahbub Alam, Md Haidar Ali, et al. (2023) Organizational Network Monitoring and Security Evaluation Using Next-Generation Firewall (NGFW). The Fourth Industrial Revolution and Beyond: Select Proceedings of IC4IR+. Singapore: Springer Nature Singapore 980: 133-147.
4. Adrian Komadina, Ivan Kovačević, Bruno Štengl, Stjepan Gro (2024) Comparative Analysis of Anomaly Detection Approaches in Firewall Logs: Integrating Light-Weight Synthesis of Security Logs and Artificially Generated Attack Detection. Sensors 24(8): 2636.
5. Malak Aljabri, Amal A. Alahmadi, Rami Mustafa A. Mohammad, Menna Aboulnour, Dorieh M Alomari, et al. (2022) Classification of firewall log data using multiclass machine learning models. Electronics 11(12): 1851.
6. Bihl, Trevor, et al. (2020) Topological data analysis for enhancing embedded analytics for enterprise cyber log analysis and forensics.
7. Ucar Erdem, Erkan Ozhan (2017) The analysis of firewall policy through machine learning and data mining. Wireless Personal Communications 96: 2891-2909.
8. Awotipe, Oluwaseun (2020) Log analysis in cyber threat detection.
9. Polpinij, Jantima, Khanista Namee (2019) Internet usage patterns mining from firewall event logs. Proceedings of the 2019 International Conference on Big Data and Education, Pp. 93-97.
10. Han, Seungwoo, et al. (2021) Optimal feature selection research for firewall log analysis using Bee Swarm Optimization with Reinforcement Learning.
11. Marques, Claudio, Silvestre Malta, João Magalhães (2021) DNS firewall based on machine learning. Future Internet 13(12): 309.
12. AL-Behadili, Hayder Naser Khraibet (2021) Decision tree for multiclass classification of firewall access. International Journal of Intelligent Engineering and Systems 14(3): 294-302.
13. Kumar, Arun, et al. (2018) Analysis of network traffic and security through log aggregation. International Journal of Computer Science and Information Security (IJCSIS) 16(6).
14. Ertam, Fatih, and Mustafa Kaya (2018) Classification of firewall log files with multiclass support vector machine." 2018 6th International symposium on digital forensic and security (ISDFS). IEEE.
15. Fazila Malik, Qazi Waqas Khan, Atif Rizwan, Rana Alnashwan, Ghada Atteia (2024) A Machine Learning-Based Framework with Enhanced Feature Selection and Resampling for Improved Intrusion Detection. Mathematics 12(12): 1799.
16. Qazi Waqas Khan, Khalid Iqbal, Rashid Ahmad, Atif Rizwan, Anam Nawaz Khan, et al. (2024) An intelligent diabetes classification and perception framework based on ensemble and deep learning method. Peer J Computer Science 10: e1914.
17. Qazi Waqas Khan, Bong Wan Kim, Rashid Ahmed, Atif Rizwan, Anam Nawaz Khan, et al. (2023) Predictive Modeling of Water Table Depth, Drilling Duration, and Soil Layer Classification Using Adaptive Ensemble Learning for Cost-Effective Percussion Water Borehole Drilling. IEEE Access.
18. Hornyák Olivér, László Barna Iantovics (2023) AdaBoost Algorithm Could Lead to Weak Results for Data with Certain Characteristics. Mathematics 11(8): 1801.
19. Devi, D Navya, K Sreenivasulu, M Janardhan (2024) Detection and Prevention of DDoS Attacks in Software-Defined Cloud Networks Using Advanced Support Vector Machine. Disruptive technologies in Computing and Communication Systems. CRC Press 46-51.
20. Fahad Ali Khan, Gang Li, Anam Nawaz Khan, Qazi Waqas Khan, Myriam Hadjouni, et al. (2023) AI-Driven Counter-Terrorism: Enhancing Global Security Through Advanced Predictive Analytics. IEEE Access 11: 135864-135879.
21. Mona Alduailij, Qazi Waqas Khan, Muhammad Tahir, Muhammad Sardaraz, Mai Alduailij, et al. (2022) Machine-learning-based DDoS attack detection using mutual information and random forest feature importance method. Symmetry 14(6): 1095.