ML-Based Detection of DDoS Attacks Using Evolutionary Algo-Rithms Optimization

Fauzia Talpur , Imtiaz Ali Korejo * , Aftab Ahmed Chandio , Ali Ghulam , Sajjad Hussain Talpur

Posted Date: 15 January 2024

doi: 10.20944/preprints202401.1099.v1

Keywords: DDoS; XGB-GA; RF-GA; SVM-GA; TOPT; Genetic Programming; Machine Learning

Article

ML-Based Detection of DDoS Attacks Using Evolutionary Algorithms Optimization

Fauzia Talpur [1], Imtiaz Ali Korejo [1,*], Aftab Ahmed Chandio [1], Ali Ghulam [2] and Sajjad Hussain Talpur [2]

[1] Institute of Mathematics & Computer Science, University of Sindh, Jamshoro 70680, Pakistan
[2] Information Technology Centre, Sindh Agriculture University, Tandojam, Sindh, Pakistan
* Correspondence: imtiaz@usindh.edu.pk

Abstract: The escalating reliance of modern society on information and communication technology has rendered it vulnerable to an array of cyber-attacks, with distributed denial-of-service (DDoS) attacks emerging as one of the most prevalent threats. This paper delves into the intricacies of DDoS attacks, which exploit compromised machines numbering in the thousands or more to disrupt data services and online commercial platforms, resulting in significant downtime and financial losses. Recognizing the gravity of this issue, various detection techniques have been explored, yet the quantitative and prior detection of DDoS attacks has seen a decline in recent methods. This research introduces an innovative approach by integrating evolutionary optimization algorithms and machine learning techniques. Specifically, the study proposes XGB-GA Optimization, RF-GA Optimization, and SVM-GA Optimization methods, employing Evolutionary Algorithms Optimization with TOPOT-Genetic Programming. Datasets pertaining to DDoS attacks were utilized to train machine learning models based on XGB, RF, and SVM algorithms, and 10-fold cross-validation was employed. The models were further optimized using evolutionary algorithms, achieving remarkable accuracy scores: 99.99% with the XGB-GA method, 99.50% with RF-GA, and 99.99% with SVM-GA. Furthermore, the study employed TPOT to identify the optimal algorithm for constructing a machine learning model, with the genetic algorithm pinpointing XGB-GA as the most effective choice. This research significantly advances the field of DDoS attack detection by presenting a robust and accurate methodology, thereby enhancing the cybersecurity landscape and fortifying digital infrastructures against these pervasive threats.

Keywords: DDoS; XGB-GA; RF-GA; SVM-GA; TOPT; genetic programming; machine learning

0. Introduction

The vast majority of the work in this area has focused on a Distributed Denial-of-Service (DDoS) assault is when numerous computers are combined as an attack platform using client/server technologies, and then the attacks are launched at one or more targets to boost the attack's potency. An example of an attack is a Distributed Denial-of-Service (DDoS) attack, where tens of thousands or even hundreds of thousands of compromised computers are utilized to target online businesses and information-providing services. This often results in substantial periods of inactivity and financial damages, as well as the denial of services to legitimate customers. The investigation of DDoS attacks is a prominent topic of research, and several methods for spotting DDoS attacks have been put forth in the literature, including evolutionary algorithms and artificial intelligence. Unfortunately, current, well-known DDoS detection techniques are losing their ability to reliably identify DDoS attacks in advance and objectively.

This strategy is centered on developing an intrusion detection system (IDS) that can discriminate between regular and attack traffic while also meeting the needs of the monitored environment. For our experimental analysis, the publicly accessible datasets KDD Cup 99 and CIC-IDS 2017 were employed. DDoS with decision tree obtains good detection and accuracy with a low false-positive rate, according to the simulation's results. The last few years have seen an increased interest in Genetic Algorithm (GA) was created by John Holland at the University of Michigan in

 doi:10.20944/preprints202401.1099.v1

1975. The GA is based on genetics and natural selection concepts (Fraser 1957, Bremermann 1958, Holland 1975). In computers, a genetic algorithm is a search strategy that is used to identify alternative solutions to optimize and search problems. A genetic algorithm uses evolutionary principles to find the best results that are close to the real ones. A genetic algorithm is an evolutionary algorithm that uses basic ideas to evaluate things in a way that is similar to how things work in nature. It has four main steps: figuring out the health of the population, reproducing, crossing over (recombination), changing a gene, and ending. In another stage called a crossover, two parents choose a good group and then mix (recombine) to make a new child.

1. Related Works

Cloud computing offers a flexible framework that enables applications to obtain the necessary resources before being executed on Virtual Machines (VMs) [1]. Pay-as-you-go pricing models are frequently used by cloud service providers, which can benefit cloud customers by lowering costs and enhancing flexibility. The wide spectrum of advancements in cloud computing technology have led to a considerable rise in cloud users and the creation of applications that can access various cloud services [2]. Through the usage of the internet and remote access to software and hardware from far-off areas, the world has transformed into a global village. The services available online represent a revolution in this century's computing industry [3]. More and more processing power is required to complete complex tasks. To complete these tasks, sophisticated and high-performance computers is required. Paying for the usage of high-performance computing hardware on a per-use basis is preferable to buying new equipment. Users are given these privileges by cloud service providers so they can employ a pay-per-use business model to access the resources and services [4]. Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) are the three basic categories into which cloud computing services are divided [5]. Users are not constrained by their location; they can use services at any time and from any location [6].

In order to deliver IoT services in accordance with user needs, a number of sensors are often put in the environment while planning an IoT system. When a user passes by a sensor in an IoT-based home system, all of the IoT services connected to that sensor are immediately active [7]. When deploying sensors in an IoT system, the optimum mapping of sensor devices to service components under the user's needs becomes a difficulty. The primary issue is how to best achieve one or more goals connected to the allocations made under the energy and distance constraints [8]. Machine learning (ML) is becoming increasingly popular in the field of medicine, particularly in the areas of diagnosis and treatment management [9]. Numerous studies have been conducted to determine how ML can increase the timeliness and precision of diagnosis [10,11]. A crucial element of all healthcare systems around the world is accurate diagnosis. A mistaken diagnosis for a significant medical illness is given to about 5% of outpatients in the US [12]. Recently, reduced-space multistream classification based on Multi-objective Evolutionary Optimization proposed by the researchers [13,14].

This paper employs neural networks for cloud resource consumption prediction with these factors in mind. Training the network weights is the main challenge in putting neural networks into practice [15]. The network's weights training is a challenging optimization challenge. For these issues, swarm and evolutionary algorithms are frequently used. Over the usage of conventional mathematical approaches, these techniques are favored [16]. In [17] proposed a meta-heuristic approaches with immigrant techniques for nurse duty roster in public hospitals in Sindh, Pakistan. The suggested model employs a hybrid Genetic Algorithm (GA) and Particle Swarm Optimization (PSO) (GA-PSO) that capitalizes on the advantages of both techniques to accomplish this goal. Impressive results have been obtained when hybrid versions of these algorithms are utilized in a number of different disciplines. Antenna array pattern synthesis [18], mining association rules [19], forecasting power consumption [20], allocating resources in cloud computing [21], and process planning [22] are a few of the applications. The simulation used by the authors of [23] has demonstrated that the hybrid version outperforms the drawbacks of the individual approaches [24].

 doi:10.20944/preprints202401.1099.v1

In a cloud computing context, processing is done by cloud servers housed in data centers that offer infrastructure, software, and platforms as Internet-based services rather than by local computers. In fact, cloud computing's objective is to combine hardware and software as a service that is available to consumers via the Internet [25, 26]. Popular cyberattacks include denial of service attacks, distributed denial of service attacks [23], remote to local attacks [27], probing attacks, user to root attacks, adversarial attacks, poisoning and evasion attacks, botnet attacks, phishing attacks[28], spamming attacks[29] and zero-day attacks [30]. There is a consensus that integrating the grasshopper optimization algorithm (GOA) with a machine learning technique called GOIDS can effectively mitigate denial of service threats. This approach focuses on creating an intrusion detection system (IDS) that can effectively differentiate between normal and malicious network traffic, while simultaneously satisfying the requirements of the monitored environment [31].

2. Materials and Methods

2.1. Datasets and Source

The proposed algorithm for datasets construction gives step-by-step directions on how to put together a dataset. During the monitoring interval events information request flow statistics are retrieved by using the OFPPortStatsRequest method. After then there is an event request and reply handler for every event. Afterwards, the OFPFlowStatsReply method is called, which is responsible for returning the flow statistics as a response to the event. Many techniques, particularly data mining techniques like clustering and data classification algorithms, were previously employed in knowledge discovery of databases (KDD) [32]. KDD deals with removing valuable data from the data source. The dataset receives its annotations in a computerized way as a result of the application of some coding logic. The programming is designed in such a way that the label column of the dataset is set to "0" when the benign traffic is running, but it is set to "1" when the malicious traffic is running. There is a "1" in the traffic label. Following the annotation of the data, we then will classify the traffic using any machine learning algorithm. The numbers 0 and 1 respectively represent the two different types of traffic, which are referred to as a) benign and b) malicious.

2.2. Proposed Novel Hybrid Method for DDoS Attack Detection using TOPT with genetics algorithm

This study used qualitative techniques to analyze computer networks' security has benefited greatly from the use of Distributed Denial-of-Service (DDoS) network attacks. An interesting side finding was that identify several attack vectors and instances of illegal software activity that firewalls may occasionally miss. Many DDoS attacks have been enhanced to classify network traffic as regular or abnormal using machine learning methods. The proposed method two phases of the new hybrid DDoS detection method in the first phase for feature selection and another phase for attacks detection are described in this work.

The employed a supervised leaning method for classification. These findings would suggest that then integrated XGB algorithm with GA, RF algorithm integrated with GA and SVM algorithm integrated with GA using with TPOT: Pipelines Optimization with Genetic Algorithms applied during the feature selection stage as shown in Figure 1. A further complication for the present hypothesis is that proposed and implemented initial scheme model framework used for building ML Pipeline 1, Pipeline2 and total number of Pipeline N as machine learning algorithms, we then integrated with genetic algorithm associated with optimal ML Pipeline.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1099.v1

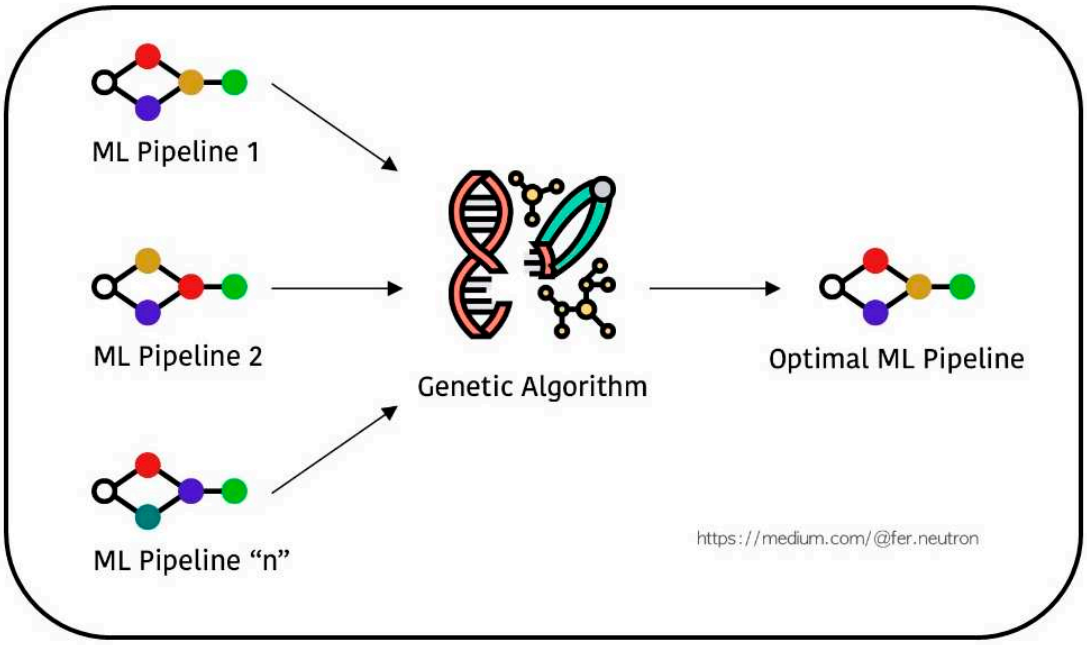

Figure 1. TPOT: Pipelines Optimization with Genetic Algorithms.

2.3. Three machine learning algorithm Integration with Genetics Algorithms

The hypothesis that conducted all analyses using XGB-GA optimization method, RF-GA optimization and SVM-GA optimization features such as multi-parent crossover and multi-parent mutation are combined in this method. An XGB-GA utilized to detect network attacks during the attack detection phase. The classifier is trained using a hybrid XGB-GA optimization and genetic programing algorithm optimization (GAO) approach to enhance performance. The proposed XGB-GA optimization method, RF-GA optimization and SVM-GA optimization based on evolutionary algorithms optimization. XGB-GA, RF-GA and SVM-GA are the name given to the proposed hybrid approach. Several findings of this study warrant further discussion, such as how well the suggested evolutionary model worked in terms of accuracy compared to seven other algorithms: ET-GA, KNN-GA, BernoulliNB-GA, GBoosting-GA, SGD-GA, MultinomialNB-GA, and LR-GA. The expected results show that the proposed XGB-GA, RF-GA, and SVM-GA methods can achieve a maximum detection accuracy of 99.00%. The dimension reduction happened when we used the NSL-KDD datasets with 42 to 16 features, and the maximum training time was only 10 seconds. The NSL-KDD dataset was used as a standard to test the attack detection methods.

2.4. The framework of the proposed DDoS diagnosis procedure

The general structure of the proposed diagnosing procedure is depicted for machine learning in Figure 1. Once we have the dataset, we may go on to the next framework in Figure 2. to see if any pre-processing is required to eliminate missing values or to replace them with suitable data for genetic algorithm. Even though we could have eliminated the faulty rows of our datasets, the opted to fill in the missing values automatically by taking the average of the remaining ones. Following this step, GA is applied on the now-clean dataset in order to determine which subset of characteristics yields the highest correlation to the targets.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1099.v1

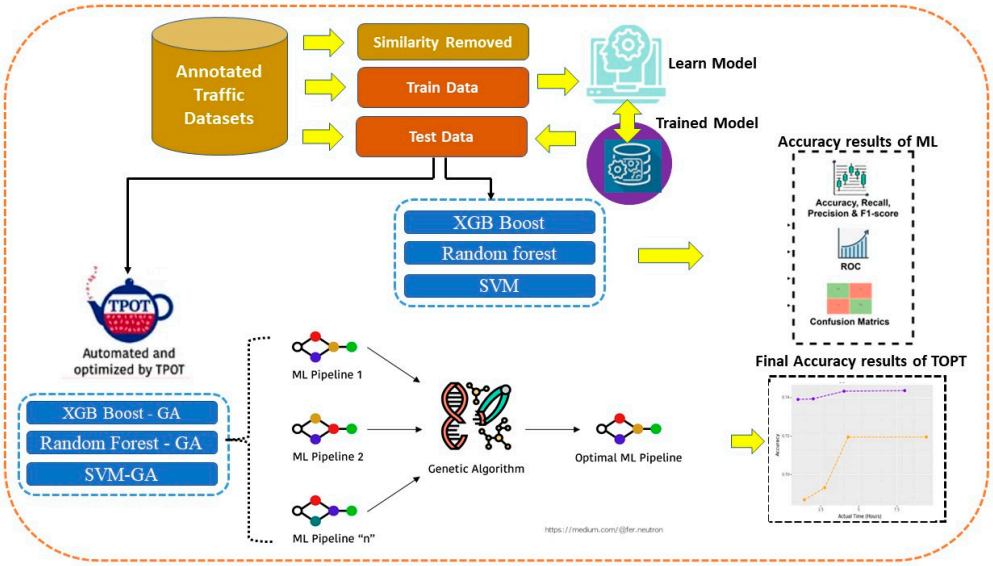

Figure 2. The framework of the proposed DDoS diagnosis procedure.

The aim is to evaluate TPOT's capabilities and decide whether or not you should incorporate it into your existing machine learning process. Future studies will have to concentrate still be needed for the foreseeable future, although automated machine learning may speed up the process of finding effective models. Genetic programming is used by TPOT, a Python Automated Machine Learning tool, to optimize machine learning pipelines.

3. Performance parameters

A lot of different Machine-Learning models have been trained and tested on our dataset. Several measures, such as Accuracy, Sensitivity, Specificity, Precision, and F1-score, were used to check how well each model worked. The true and false values of the classifier are displayed in a 22 matrix, which is the definition of the error matrix in a binary classification problem. Below is an explanation of the matrix's four values which can be somewhat perplexing at first glance.

- True Positive tp: When both the model's forecast and the actual values in the dataset are positive, we say that a value is true positive, or tp. meaning the classifier accurately differentiates between good and bad traffic.
- True negative tn: When both the model's forecast and the actual values in the dataset are negative, we say that b) is a true negative tn. i.e. it is the circumstance where the traffic is accurately categorized as malicious.
- False Positive fp: False positive is the error category where the model prediction is positive and actual values in the dataset is negative. i.e. \sit is the circumstance where the traffic is wrongly classed as innocuous.
- False Negative fn: A false negative is a form of error in which the actual values in the dataset contradict the prediction of the model. i.e., it is the circumstance where the traffic is wrongly categorized as harmful.
- As a performance metric, accuracy may be written as a fraction with the sum of correct answers (positive and negative) in the numerator and the sum of incorrect answers (positive and negative) in the denominator.

$$Sensitivity = \frac{TP}{TP+FN} \tag{1}$$

$$Specificity = \frac{TN}{TN+FP} \tag{2}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1099.v1

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$\text{Precision}(p) = \frac{TP}{TP + FP} \tag{4}$$

$$F1score = \frac{2 * R * P}{R + P} \tag{5}$$

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{6}$$

4. Results

4.1. Three machine learning classification results

In Figure 3. A possible interpretation of this finding is that compare the SVM-RF hybrid model's accuracy to that of many other well-defined machine learning models. Exploration of the data is performed to gain comprehension of the datasets, the distribution of normal and harmful traffic within the dataset, and the number of instances in each type of traffic. The tabular format below provides a concise summary of the analyzed dataset. A further complication for the present hypothesis is that how much good and bad traffic makes up each type of traffic. In Table 1, A possible interpretation of this finding is that can see how much good and bad traffic makes up each type of traffic. To better understand the dataset some summary information is provided.

Table 1. Traffic category of each traffic instance.

Traffic Class	Benign	Malicious
ICMP	24957	16364
TCP	18897	10539
UDP	22772	10816

4.2. SVM

Once the dimensions have been minimized, SVC can be used to fit the data, as illustrated in the image. As may be seen in Fig. 3, This demonstrates that with repeated training on the dataset, the model becomes more accurate at classifying traffic. The model's inability to appropriately capture the linear relationship between the features led to inaccurate predictions of the class labels, hence it does not provide adequate results. It also has extensive citations in the academic literature.

4.3. Random Forest" (RF)

Classifier known as "Random Forest" (RF) that makes use of many decision trees to reach a conclusion. It is possible for other decision trees to correct for an incorrect one. Each decision tree outputs a categorization result, with the highest scores being weighted toward the proposed ultimate score. This demonstrates that Rf is the superior classifier.

4.4. XGBoost

Ensemble classifiers (ECs) are a type of classifier that combines the results of multiple classifiers into a single one. Classifiers like XGBoost, Random Forest, and SVM are used. The classifier has a 99.9% accuracy, which is significantly higher than the performance of separate classifiers.

Combining the strengths of SVC and RF, or "Support Vector and Random Forest," creates a powerful new classification method. This classifier achieves the best results on our dataset when two machine learning methods are combined. In Table 2. this study used qualitative in order to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1099.v1

determine the aforementioned efficiency metrics, the confusion matrix is frequently employed. The analysis was based on used the dataset outlined in Table 2, the results analysis consists of calculated accuracy, precision, recall, and F1-score to evaluate the system's overall performance.

Table 2. Performance Measures of different Algorithms.

Algorithms	accuracy	precision	recall	f1-score
SVM	72.00%	0.71	0.83	0.76
Random forest	98.00%	0.98	0.99	0.98
XGBoost	98.08%	0.99	0.99	0.98

4.5. Receiver Operating Characteristic ROC(AUC) training performance

This comparison analysis is further expanded to analyze the training performance of machine learning models through the use of k-fold cross validation, Area Under Curve (AUC) analysis of the Receiver Operating Characteristic (ROC) curves. The Receiver Operating Characteristic (ROC) curve is a tool that can be used to evaluate the performance of a classification model by taking into consideration the False Positive Rate (FPR) and the True Positive Rate (TPR). In addition, Figure 3, are utilized to illustrate the ROC(AUC) graphs of RANDOM FOREST, DECISION TREE, NAIVE BAYES, and SVM, respectively.

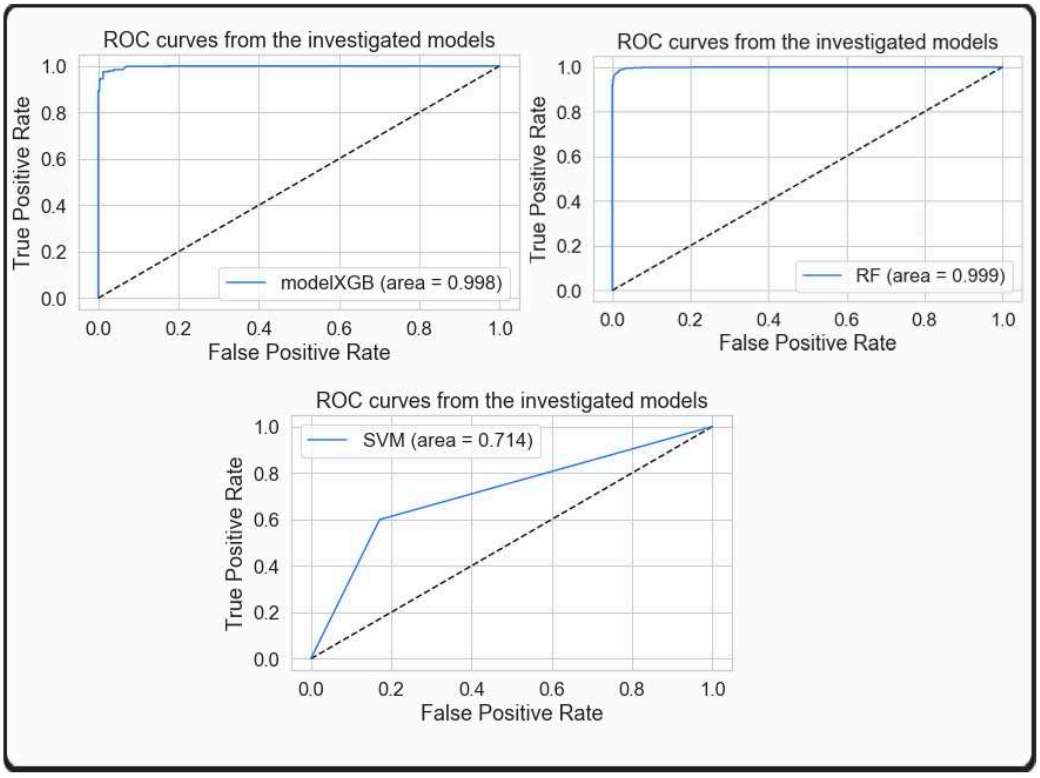

Figure 3. ROC(AUC) performance of a classification model.

4.6. Performing accuracy tests using a variety of methodologies for fivefold cross validation

It has been explained in the part titled "accuracy performance" that various different machine learning strategies were utilized in order to analyze the dataset. The dataset was subjected to five rounds of cross-validation, one of which was performed with each of these methods. Figure 4. illustrate how the accuracy results converge after being subjected to five iterations of cross-validation. It can be deduced from the results of the cross-validation that the Logistic Regression Decision Tree and KNN models performed the best in terms of accuracy throughout training and testing. When it came to training, Decision Tree performed in a manner that was approximatively

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1099.v1

comparable to Random Forest and Naïve Bayes. In the instance of the testing, the KNN performed poorly at initially and had the highest number of deviations compared to the other methods.

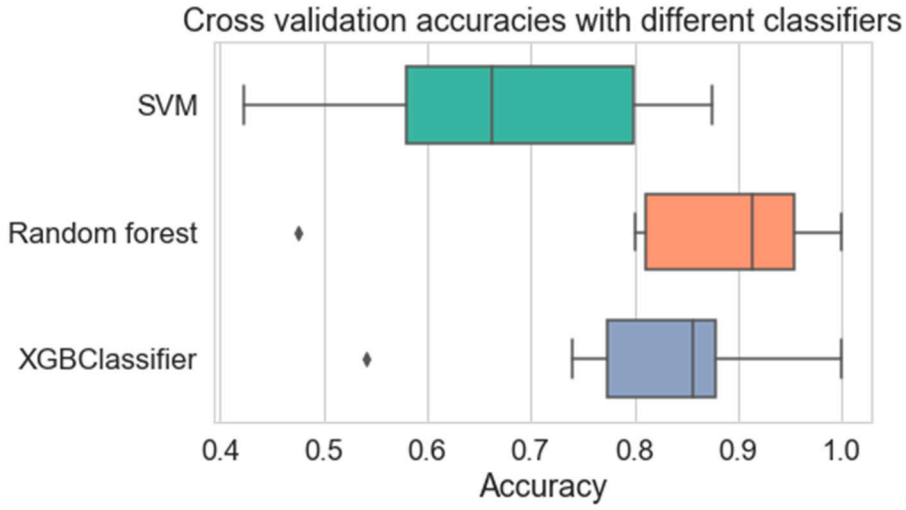

Figure 4. Performing accuracy tests using based different classification model.

5. Three machine learning algorithm optimizations with genetic algorithms results

A more advanced technique to machine learning is used in automated machine learning (AutoML). TPOT for a short while and it will discover a passably efficient pipeline our DDoS datasets dataset. To enable TPOT to fully search the pipeline space DDoS dataset, it is frequently beneficial to run several instances of the program in parallel for a considerable amount of time (hours to days).

AutoML algorithms involve more than just fitting a single model to the DDoS dataset; we have calculated a variety of machine learning algorithms (XGB-GA Optimization, RF-GA Optimization and SVM-GA) in a pipeline that includes a number of preprocessing steps (like feature selection, scaling, PCA, missing value imputation), the hyperparameters for each model preprocessing step and then we have deployed 5-fold validation method as 5 Generation cross-validation with using of pipeline arrangement options as shown Table 3. We improved obtained best pipeline test accuracy 1.000 using with genetics algorithm. This demonstrates how a straightforward genetic algorithm implementation can enhance the performance measure in XGBoost. Two criteria were used to assess the performance of the pipelines: the trained pipeline's classification accuracy and the training pipeline's elapsed time as shown in Table 4.

Table 3. Best pipeline: XGB-GA Optimization Accuracy Score.

Classifiers	MAE	MSE	R^2
Gradient Boost	4.7917	4.7917	0.9997

Table 4. XGB-GA Best metric performance.

Classifiers	GA Generations	Best internal CV score	GA Optimization Best Accuracy Score
Gradient Boost	Generation 1	Current best internal CV score:	0.9999
	Generation 2	Current best internal CV score:	1.0
	Generation 3	Current best internal CV score:	1.0
	Generation 4	Current best internal CV score:	1.0
	Generation 5	Current best internal CV score:	1.0
		Best pipeline test accuracy:	**1.000**
		Accuracy:	**0.9999**

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1099.v1

It is generally agreed that computed MSE error rate in the missing numeric column values with the column's most frequent value. Current research seems to indicate that divide the columns into categories and employ various imputation techniques according to whether the data were nominal, ordinal, or interval. Because not every column had a most frequent value, missing values in string columns were filled in with a "missing" label before they were ordinally encoded. Then, each feature would have an additional dimension created by TPOT's single Gradient Boosting framework process, indicating that there was a missing value in the data for that feature. By default, TPOT Gradient Boosting framework employs mean squared error scoring.

The runs concluded rather fast, with distinct winning pipelines identified each time. The Root Mean Squared Error (RMSE) is obtained by taking the square root of the scores as shown in Table 5. Three methods for Hyper Parameter Optimization (HPO) To improve the performance of the XGB classifier model, RF classifier model and SVM classifier model and Genetic Programming (TPOT Classifier) were suggested. The two models, XG Boost, Random Forest and SMV were evaluated against the results of previous research. All four did a good job classifying the XGB-GA Optimization DDoS attack traffic, and we can both identify benign traffic with a 100% accuracy rate. unfortunately, compared to the other seven optimization method using with TPOT classifiers, XGB-GA Optimization performed well. XGB-GA Optimization accurately identify the DDoS attack traffic. The results of the best pipeline test accuracy score can also be confirmed by looking at the Table 8.

Table 5. Best pipeline: XGB-GA MSE Error rate.

Classifiers	Precision	Recall	F1-Score	Accuracy
Gradient Boost	1.00	1.00	1.00	0.9999
	1.00	1.00	1.00	

5.1. RF-GA Optimization with genetic algorithms results

In order to develop the model, we conducted all analyses using two algorithms random forest-GA optimization and SVM-GA Optimization. The analysis was based on contrast each algorithm's accuracy score. This helps us determine which is superior. This study used qualitative utilization of TPOT to create a machine learning model in the following section. TPOT first selects the best classification algorithm by combining all of the available ones. The utilization of genetic algorithm optimization found the who scored the highest on accuracy is then selected. The DDoS attack traffic dataset used to train our model. Based on the input features, the model categorizes the different types of attacks as shown in Table 6.

Table 6. Best pipeline: RF-GA Optimization Accuracy Score.

Classifiers	GA Generations	Best internal CV score	Best Pipeline Test Accuracy Score
RF-GA	Generation 1	Current best internal CV score:	0.9981
	Generation 2	Current best internal CV score:	0.9988
	Generation 3	Current best internal CV score:	0.9983
	Generation 4	Current best internal CV score:	0.9988
	Generation 5	Current best internal CV score:	**0.9988**
Best pipeline test accuracy:			**0.9988**

Our model is constructed using the Random Forest Classifier, which is the second approach as integrated with genetic algorithm RF-GA Optimization. The second approach RF-GA Optimization have employed. Next, were computed using contrast the accuracy rating. After developing our model, we will utilize TPOT to aggregate all of the methods and identify the optimal one. The second technique we employed to model the data was Random Forest. The accuracy scores of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1099.v1

two algorithms can then be compared. Accuracy ratings from RF-GA Optimization range from 0.9999 to 0.9960. Evidently, RF-GA Optimization is superior then SVM-GA optimization.

5.2. SVM-GA optimization with genetic algorithms results

SVM-GA optimization is the option for a researchers would select while creating the model. Nevertheless, since we have only compared two algorithms, this one could not be the best. Building models with various methods is a laborious procedure. TPOT is therefore the ideal option when working with many algorithms. TPOT finds the optimal classification method by combining all existing ones. As a result, it saves a ton of time by automating the genetic programming model building process and eliminating the need to manually compare every viable algorithm. There SVM-GA optimization is a same procedure for optimization. To identify the ideal pipeline, TPOT will iterate five times. Beneficial since it automates the entire procedure, saving the users time. In this process of optimization, TPOT applies the theory of genetic programming.

It eventually determines the optimal algorithm as a result. We can also determine the precise parameters needed to accomplish this optimization with the aid of TPOT. We began by getting our DDoS attack traffic dataset ready. Then, we employed two techniques to create a model utilizing this dataset. To determine which algorithm was superior, we have compared the two algorithms based on genetic programing. SVM-GA Optimization best pipeline test accuracy score 9960, RF-GA Optimization best pipeline test accuracy score 0.9999. RF-GA Optimization proved to be the most effective algorithm through genetic programming which is better than SVM-GA Optimization as shown Table 7.

Table 7. Best pipeline: SVM-GA Optimization Accuracy Score.

Classifiers	GA Generations	Best internal CV score	Best Pipeline Test Accuracy Score
SVM-GA	Generation 1	Current Pareto front scores:	0.9739
	Generation 2	Current Pareto front scores:	0.9835
SVM-GA	Generation 3	Current Pareto front scores:	0.9925
	Generation 4	Current Pareto front scores:	0.9925
	Generation 5	Current Pareto front scores:	**0.9940**
Best pipeline test accuracy:			**0.9960**

5.3. Proposed three TOPOT-Classifiers with other 7 seven GA optimization models results

All four did a good job classifying the XGB-GA Optimization DDoS attack traffic, and we can both identify benign traffic with a 100% accuracy rate. unfortunately, compared to the other seven optimization method using with TPOT classifiers, XGB-GA Optimization performed well. XGB-GA Optimization accurately identify the DDoS attack traffic. The results of the best pipeline test accuracy score can also be confirmed by looking at the Table 8. ML algorithms were used to categorize the DDoS attack into several classes, and each category was then identified and verified according to different standards. A thorough examination of several GA Optimization was done for the purpose of identifying DDoS multiclass cyberthreats, with XGB-GA Optimization method, RF-GA Optimization and SVM-GA Optimization based on Evolutionary Algorithms Optimization using with TOPOT- Genetic programming reliability index of SVM-GA Optimization best pipeline test accuracy score 0.9960, RF-GA Optimization best pipeline test accuracy score 0.9950 and XGB-GA Optimization method best pipeline test accuracy score 1.000 accomplish the goal as shown Table 6. comparison results many more DDoS attack types may be addressed for categorization and prediction in the future.

A recent line of research has focused on best pipeline test accuracy 1.000% best pipeline accuracy score based on XGB-GA optimization method, 0.9950% best pipeline accuracy score based on RF-GA optimization and 0.9999% best pipeline accuracy score based on SVM-GA optimization

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 15 January 2024 doi:10.20944/preprints202401.1099.v1

method. Finally, we used TPOT to find the best algorithm to use when building a machine learning model. Through genetic algorithm, the best algorithm was XGB-GA algorithm.

6. Comparative analysis with existing results

The present study employed a DDOS attack dataset to compare the paper's proposed technique against previous research on DDOS attack detection in order to evaluate it (see Table 9). The best standard that is currently available is found to be 96% and TOPT Best pipeline test accuracy 1.000 It is evident from the data in Table that the study's suggested model has the highest accuracy. The EAs methodological paradigm that this study suggests is quite beneficial for identifying these attacks early on.

Table 8. TOPOT-Classifiers Optimization Comparison.

ML-GA Classifiers	5 iterations/5-fold CV	Best Pipeline Test Accuracy Score
Extra Trees Classifier	Internal cv score	0.8123
K-Neighbors Classifier	Internal cv score	0.8158
Bernoulli NB	Internal cv score	0.7322
GBoosting Classifier	Internal cv score	0.9910
SGD Classifier	Internal cv score	0.5283
Multinomial NB	Internal cv score	0.5307
Logistic Regression	Internal cv score	0.7151
SVM-GA Optimization	Internal cv score	**0.9940**
Best pipeline test accuracy	Internal cv score	**0.9960**
RF-GA Optimization	Internal cv score	**0.9988**
Best pipeline test accuracy	Internal cv score	**0.9950**
Proposed XGB-GA	Accuracy:	**0.9999**
	Best pipeline test accuracy:	**1.000**

A comparison is made between the suggested approach in the paper (see Table 9) and the previous research in the field of DDOS attack detection using emulated datasets. 96% is confirmed to be the highest benchmark result currently in use. As can be observed from the Table, our suggested model gets the highest accuracy of accuracy 1.000. For the suggested model, the EAs method is important for attack detection. With a shorter training time, this model turned out to be the highest performing model for our dataset.

Table 9. Comparison Results of traffic classification using various Simulated SDN Datasets.

S. No	Authors	Testing Accuracy
1	Meti et al., 2017 [33]	80%
2	Da Silva et al., 2016 [34]	88.7%
3	Perez-Díaz et al. [35]	95%
4	Ye et al., 2018 [36]	95.24%
5	Ko et al. [37]	96%
6	Han et al., 2018 [38]	96%
7	MyintOo et al., 2019 [39]	97%
8	Auhoja [40]	98.8%
9	Proposed XGB-GA Optimization	**99.00%**
10	Proposed TOPT Best pipeline test accuracy:	**1.000%**

7. Discussion

The aim is to evaluate TPOT's capabilities and decide whether or not you should incorporate it into your existing machine learning process. Besides that, TPOT's will choose the most applicable features from the original IDS dataset that can aid in distinguishing typical low-speed DDoS attacks

 doi:10.20944/preprints202401.1099.v1

and these features are then passed to classifiers such as support vector machine, decision tree, nave Bayes, and multilayer perceptron to identify the type of attack. The simulation results will show that evolutionary algorithms (EAs) and ML classification method will achieves good detection and accuracy with a low false-positive rate. To sum up TPOT- DDoS represents a promising advancement in automating the creation of machine learning workflows for cyber security through the use of evolutionary algorithms. The research design involved the field of automated machine learning is well-suited for evolutionary algorithms (EAs), and specific instruments like TPOT-DDoS accentuate the benefits of EAs by demonstrating how simple an EA solution. A dataset's features can be used by machine learning algorithms to learn new things. For the purpose of this study was measured with refer to it as machine learning model training. The traffic can be divided into classifications by the trained model: malicious and benign. To classify the traffic, the trained model can also be used in real-time.

8. Conclusion

The present study employed a quantitative proposed XGB-GA Optimization method, RF-GA Optimization and SVM-GA Optimization based on Evolutionary Algorithms Optimization using with TOPOT- Genetic programming to find the highest score as the model accuracy. The networking architecture that software has defied is called software-defied networking, or SDN. The controller, which remotely guides the traffic between the hosts, centrally controls network traffic. Even with such adaptable network traffic management, the network is still vulnerable to a number of threats. In this research, the authors develop an SDN dataset and use machine learning methods to distinguish between traffic from DDOS attacks and benign traffic. To deal with uncertainty in cloud computing environments, it is important to be able to predict how cloud resources will be used. In cloud computing, users are given access to their applications from anywhere in the world via the Internet. Internet-based technology and online services are very important in the world of technology today. Services on the Internet are now a part of everyone's daily life. This kind of service dependency has led to a new kind of change and has opened the door to attacks on network services. In DDoS attacks, multiple DoS attacks are launched against the victim (the destination server) at the same time by multiple infected systems acting as attack agents. This makes it so that a specific service isn't available by flooding the service provider's resources with false requests, which is a huge risk for the network. Because of how hard it is to identify DDoS attacks with the current countermeasures, many new techniques are needed to find and stop DDoS attacks more effectively. This approach is performed in two steps: firstly, initially the features are selected through XGB-GA, RF-GA and SVM-GA and then selected features are passed to the different classifiers XGB-GA, RF-GA and SVM-GA to classify DDoS attack. Future research could focus on evaluating neural network predictors in other areas of cloud computing, such as predicting other resources like disc usage, cost-effectiveness, network, and lowering energy use for green computing. It would be helpful to support the proposed evolutionary neural network approach by working on other multivariate datasets of resource usage. The investigation of DDoS attacks is a prominent topic of research, and several methods for spotting DDoS attacks have been put forth in the literature, including evolutionary algorithms and artificial intelligence. Regrettably, current, well-known DDoS detection techniques are losing their ability to reliably identify DDoS attacks in advance and objectively. The present study employed a quantitative research method to examine and obtained 1.000% best pipeline accuracy score based on XGB-GA optimization method, 0.9950% best pipeline accuracy score based on RF-GA optimization and 0.9999% best pipeline accuracy score based on SVM-GA optimization method. Finally, this study used a combination of qualitative and quantitative analysis tools TPOT to find the best algorithm to use when building a machine learning model. Through genetic algorithm, the best algorithm was XGB-GA algorithm. We used ML algorithms and integrated with TOPT-GA algorithm. And first time used DDOS attack detection based on ML-GA and received optimized results.

Author Contributions: For Conceptualization, F.T. and I.A.; methodology, F.T.; software, F.T.; validation, A.A., A.G. and S.H.; formal analysis, F.T.; investigation, A.G.; resources, F.T.; data curation, F.T.; writing—

original draft preparation, S.H.; writing—review and editing, A.G.; visualization, I.A.; supervision, I.A.; All authors have read and agreed to the published version of the manuscript.

Funding: Please add: This research received no external funding.

Institutional Review Board Statement: "Not applicable".

Informed Consent Statement: "Not applicable".

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bhutto, A., Chandio, A. A., Luhano, K. K., & Korejo, I. A. Analysis of Energy and Network Cost Effectiveness of Scheduling Strategies in Datacentre. Cybernetics and Information Technologies, 2023. 23(3), 56-69.
2. Chandio, A. A., Korejo, M. S., Korejo, I. A., & Chandio, M. S. To investigate classical scheduling schemes with power management in IaaS cloud environment for HPC workloads. In 2017 IEEE 15th Student Conference on Research and Development (SCOReD). 2017. (pp. 121-126). IEEE.
3. Wyld, D. C. The cloudy future of government IT: Cloud computing and the public sector around the world. International Journal of Web & Semantic Technology. 2010. 1(1), 1-20.
4. Muteeh, A., Sardaraz, M., & Tahir, M. MrLBA: multi-resource load balancing algorithm for cloud computing using ant colony optimization. Cluster Computing, 2021. 24(4), 3135-3145.
5. Maryam, K., Sardaraz, M., & Tahir, M. Evolutionary algorithms in cloud computing from the perspective of energy consumption: A review. In 2018 14th international conference on emerging technologies (ICET), 2018. (pp. 1-6). IEEE.
6. Ganesh Kumar, G., & Vivekanandan, P. Energy efficient scheduling for cloud data centers using heuristic based migration. Cluster Computing, 2019. 22, 14073-14080.
7. Younas, I., & Naeem, A. Optimization of sensor selection problem in IoT systems using opposition-based learning in many-objective evolutionary algorithms. Computers & Electrical Engineering, 2022. 97, 107625.
8. Gubbi, J., Buyya, R., Marusic, S., & Palaniswami, M. Internet of Things (IoT): A vision, architectural elements, and future directions. Future generation computer systems, 2013. 29(7), 1645-1660.
9. Al Bataineh, A., & Manacek, S. MLP-PSO hybrid algorithm for heart disease prediction. Journal of Personalized Medicine, 2022. 12(8), 1208.
10. Samieinasab, M., Torabzadeh, S. A., Behnam, A., Aghsami, A., & Jolai, F. Meta-Health Stack: A new approach for breast cancer prediction. Healthcare Analytics, 2022. 2, 100010.
11. Jiao, B., Guo, Y., Yang, S., Pu, J., & Gong, D. Reduced-space Multistream Classification based on Multi-objective Evolutionary Optimization. IEEE Transactions on Evolutionary Computation, 2022.
12. Hameed, B. Z., Prerepa, G., Patil, V., Shekhar, P., Zahid Raza, S., Karimi, H., ... & Somani, B. K. Engineering and clinical use of artificial intelligence (AI) with machine learning and data science advancements: Radiology leading the way for future. Therapeutic Advances in Urology, 2021. 13, 17562872211044880.
13. Tuli, S., Ilager, S., Ramamohanarao, K., & Buyya, R. Dynamic scheduling for stochastic edge-cloud computing environments using a3c learning and residual recurrent neural networks. IEEE transactions on mobile computing, 2020. 21(3), 940-954.
14. Hu, C., Zeng, S., & Li, C. An uncertainty measure for prediction of non-Gaussian process surrogates. Evolutionary Computation, 2023. 31(1), 53-71.
15. Zelinka, I. A survey on evolutionary algorithms dynamics and its complexity–Mutual relations, past, present and future. Swarm and Evolutionary Computation, 2015. 25, 2-14.
16. Casalino, L., Masseni, F., & Pastrone, D. Robust Design Approaches for Hybrid Rocket Upper Stage. Journal of Aerospace Engineering, 2019. 32(6), 04019087.
17. Jatoi, W. M., Korejo, I. A., Chandio, A. A., Brohi, K., & Koondhar, Y. M. Meta-heuristic algorithms with immigrant techniques for nurse duty roster in public hospitals in Sindh, Pakistan, 2020.
18. Dong, D., Ye, Z., Cao, Y., Xie, S., Wang, F., & Ming, W. An improved association rule mining algorithm based on ant lion optimizer algorithm and FP-growth. In 2019 10th IEEE international conference on intelligent data acquisition and advanced computing systems: technology and applications (IDAACS), 2019. (Vol. 1, pp. 458-463). IEEE.
19. Ahmad, A. S., Hassan, M. Y., Abdullah, M. P., Rahman, H. A., Hussin, F., Abdullah, H., & Saidur, R. A review on applications of ANN and SVM for building electrical energy consumption forecasting. Renewable and Sustainable Energy Reviews, 2014. 33, 102-109.
20. Madni, S. H. H., Latiff, M. S. A., Coulibaly, Y., & Abdulhamid, S. I. M. Recent advancements in resource allocation techniques for cloud computing environment: a systematic review. cluster computing, 2017. 20, 2489-2533.

21. Wang, L. Machine availability monitoring and machining process planning towards Cloud manufacturing. CIRP Journal of Manufacturing Science and Technology, 2013. 6(4), 263-273.
22. Løken, E. Use of multicriteria decision analysis methods for energy planning problems. Renewable and sustainable energy reviews, 2007, 11(7), 1584-1595.
23. Xia, W., & Wu, Z. An effective hybrid optimization approach for multi-objective flexible job-shop scheduling problems. Computers & industrial engineering, 2005. 48(2), 409-425.
24. Aslanpour, M. S., Ghobaei-Arani, M., & Toosi, A. N. Auto-scaling web applications in clouds: A cost-aware approach. Journal of Network and Computer Applications, 2017. 95, 26-41.
25. Buyya, R., Broberg, J., & Goscinski, A. M. (Eds.). Cloud computing: Principles and paradigms. John Wiley & Sons 2010.
26. Khalaf, B. A., Mostafa, S. A., Mustapha, A., Mohammed, M. A., & Abduallah, W. M. Comprehensive review of artificial intelligence and statistical approaches in distributed denial of service attack and defense methods. IEEE Access, 2019. 7, 51691-51713.
27. Dixit, P., & Silakari, S. Deep learning algorithms for cybersecurity applications: A technological and status review. Computer Science Review, 2021. 39, 100317.
28. Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. A comprehensive survey of AI-enabled phishing attacks detection techniques. Telecommunication Systems, 2021. 76, 139-154.
29. Mohammed, M. A., Gunasekaran, S. S., Mostafa, S. A., Mustafa, A., & Abd Ghani, M. K. Implementing an agent-based multi-natural language anti-spam model. In 2018 International symposium on agent, multi-agent systems and robotics (ISAMSR). 2018.(pp. 1-5). IEEE.
30. Aburomman, A. A., & Reaz, M. B. I. A survey of intrusion detection systems based on ensemble and hybrid classifiers. Computers & security, 2017. 65, 135-152.
31. Dwivedi, S., Vardhan, M., & Tripathi, S. Defense against distributed DoS attack detection by using intelligent evolutionary algorithm. International Journal of Computers and Applications, 2022. 44(3), 219-229.
32. Natarajan, S., Mgen. https://ryu.readthedocs.io/en/latest/ryu_app_api.html, 2014.
33. Kumar, P. A. R., & Selvakumar, S. Detection of distributed denial of service attacks using an ensemble of adaptive and hybrid neuro-fuzzy systems. Computer Communications, 2013. 36(3), 303-319.
34. Da Silva, A.S., Wickboldt, J.A., Granville, L.Z., Schaeffer-Filho, A., Atlantic: a framework for anomaly traffi detection, classifiation, and mitigation in sdn. In: NOMS IEEE/IFIP Network Operations and Management Symposium. IEEE, 2016. pp. 27–35.
35. Perez-Diaz, J. A., Valdovinos, I. A., Choo, K. K. R., & Zhu, D. A flexible SDN-based architecture for identifying and mitigating low-rate DDoS attacks using machine learning. IEEE Access, 2020. 8, 155859-155872.
36. Ye, J., Cheng, X., Zhu, J., Feng, L., & Song, L. A DDoS attack detection method based on SVM in software defined network. Security and Communication Networks, 2018.
37. Ko, I., Chambers, D., & Barrett, E. Self-supervised network traffic management for DDoS mitigation within the ISP domain. Future Generation Computer Systems, 2020. 112, 524-533.
38. Han, B., Yang, X., Sun, Z., Huang, J., & Su, J. OverWatch: A cross-plane DDoS attack defense framework with collaborative intelligence in SDN. Security and Communication Networks, 2018, 1-15.
39. Myint Oo, M., Kamolphiwong, S., Kamolphiwong, T., & Vasupongayya, S. Advanced support vector machine-(ASVM-) based detection for distributed denial of service (DDoS) attack on software defined networking (SDN). Journal of Computer Networks and Communications, 2019.
40. Ahuja, N., Singal, G., Mukhopadhyay, D., & Kumar, N. Automated DDOS attack detection in software defined networking. Journal of Network and Computer Applications, 2021. 187, 103108.