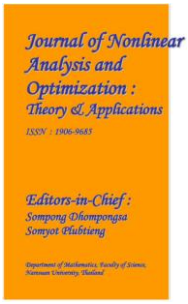
Journal of **N**onlinear **A**nalysis and **O**ptimization
Vol. 14, Issue. 2, No. 4: 2023
ISSN : **1906-9685**



CLOUD-BASED ANOMALY DETECTION WITH RXGB ALGORITHM

Dr K.Maheswari, Assistant Professor, Department of Computer Science, Anna Adarsh College for Women, Anna Nagar, Chennai 40 k.maheswari@annaadarsh.edu.in

Abstract:
In cloud computing, users have access to shared network resources like storage space and processing power whenever they need them, without the need for onsite or direct administration. Clients have access to a single online platform via CC, which is a network of both public and private data centers. It only recently materialized. A new computing paradigm known as "edge computing" is quickly gaining traction with the goal of moving computation and data storage closer to end users in an effort to save transmission capacity and decrease latency. Mobile CC (MCC) delivers applications to mobile devices using distributed computing. Association acknowledgment and client vulnerability are two security concerns with CC and edge computing that slow down the adoption of new computing paradigms. Here, we present the RXGB (robust XGBoost) anomaly detection model for distributed denial of service (DDoS) assaults. Computer algorithms that become better with use are the focus of machine learning (ML). An innovative supervised machine learning methodology called RXGB (Robust XGBoost) that makes use of the network nodes' historical knowledge and a custom-built best-effort iterative method are both included into the proposed system to enhance the accuracy of seldom detected assaults. The machine learning method generates the classifier that differentiates between the assaults being investigated. In a secure database, the system stores these options. Maximum number of reliably detectable classes and total detection accuracy are both improved by the proposed approach, even for classes with very few training entries.

Keyword: Cloud computing, XGBoost, Attack Detection.

INTRODUCTION

Using a shared pool of resources including network bandwidth, memory, processing power, and user applications is what's known as "cloud computing," a kind of Internet-based computing. With less infrastructure expense and less maintenance required, these services may be quickly and on-demand provided to end users via the Internet. Three distinct models describe the cloud computing services: Platform as a service (PaaS), infrastructure as a service (IaaS), and software as a service (SaaS) are all forms of virtualization. Additionally, it is flexible enough to be implemented as a private, public, community, or hybrid cloud. The fact that not all businesses are ready to commit fully to cloud computing is now one of the technology's major problems. Distributed denial of service attacks are a nasty way to mess with cloud servers.The current state of DDOS attack defenses requires the ability to identify fraudulent or legal data packets. These approaches may be broadly classified as signature-based or anomaly-based. Using attack signatures that have already been produced and saved in a database is the basis of the signature-based attack detection method.A cybercriminal may update the database with the definitions of a new attack based on the time since its launch, allowing them to elude detection [2].

The Internet has given many individuals a way to live more comfortably, and everyone uses it every day. The expansion of network traffic for many technologies, including the IoT, smart grids, and 5G communications, has undergone a significant transformation. Because of this, many are very worried about the Internet's security due to the unsecure communication methods [3]. However, it also made it

possible to steal other people's resources and utilize them for one's own ends. The primary function of the Internet is to facilitate communication between individuals located in different parts of the globe. Therefore, it has also made it easier for criminals to get access to the system and steal data, whether that access is permitted or not. Cybersecurity refers to the phenomena of safeguarding systems against these types of illicit activities. The aforementioned assaults have skyrocketed in frequency, and data theft has resulted from individuals sharing personal information carelessly without thinking about the potential consequences [4].In 2014, the US announced a research that found many vulnerabilities. Also reported in 2007 were cyberattacks from Russia, the severity of which varied. In order to create a more efficient Intrusion Detection System (IDS), researchers have recently integrated study regions, increased the size of datasets, implemented dynamic workspaces, and used more broad sampling. In order to differentiate between legitimate and harmful intrusions, an IDS examines and categorizes several aspects of the data flow of traffic. The cyber defense infrastructure played a part in the vulnerability assessment that will inform decisions in the future. The use of such a method allows for the systematic assessment and management of workplace risk [5].

Intrusion detection systems (IDS) are able to identify and thwart both known and unknown attacks with the use of machine learning models like XGBoost. Algorithms that can learn from past data and adapt to new patterns continually make this possible. Furthermore, machine learning has the potential to enhance data sharing protocols by adapting transmission and encryption strategies to different data types and future threat landscapes. As a result, measures for safe data exchange are more powerful.As a detection algorithm, we use the XGBoost technique. Sort the network flows into hostile and benign categories according to certain properties of the traffic [6] [7]. To start, the XGBoost classifier is a boosting classifier that iteratively builds a high-accuracy, low-false-positive model by combining hundreds of tree models with lower classification accuracy. Secondly, the algorithm is accelerated using XGBoost using two methods.For the purpose of detecting distributed denial of service attacks, we presented the RXGB (robust XGBoost) model here.

RELATED WORKS

A Distributed Denial of Service (DDoS) attack is when an attacker deliberately tries to use up all of a target node's resources.When dealing with large amounts of data, diverse formats, inconsistent flows, filtering out irrelevant information, connecting and transforming data, and data coming from several sources at varying speeds, traditional approaches may run into problems. It is in this part that we go into the various approaches to intrusion detection systems.

An method based on VFDT was proposed by Rabia Latif et al. [8]. One of their key innovations is an improved Very Fast Decision Tree (EVFDT) method for classifying attacks; this approach is different from others in the market with regard to tree size and accuracy. The synthetic dataset that is created by performing a LEACH protocol is used to test the performance of EVFDT algorithms. For a distributed denial of service (DDoS) attack, code is developed. The experimental results demonstrate that the suggested approach can identify an assault with a low false alarm rate (1.1%), a high classification accuracy (96.5%), and little memory cost. The victim node is where the suggested model is put into action in a real-time network setting.

According to Rabia Latif et al. [9], data mining methods have been an invaluable resource for several fields of study. Using typical data mining approaches in sensor WBAN is time-consuming and complicated since they rely on structured, fixed computer units, store data in databases, and need frequent updates and retraining. Instead of using the old data mining methods used by WBAN, VFDT takes use of the massive amounts of incoming data streams. Because of its small size and lack of storage requirements, VFDT is an excellent choice for WBANs with limited resources.

Using z-tests, Joffrey L. Leevy et al. [10] found that LightGBM is a better classifier of CSE-CIC-IDS2018 data than XGBoost, even without using Destination Port as a feature for LightGBM. This is true regardless of the settings of XGBoost and LightGBM that they utilize. Also included in the summary of their research topics was a proposal to combine their ensemble feature selection approach with LightGBM to decrease computing needs and discover anomalies (attacks) in the 2018 dataset.

A solution based on machine learning was suggested by Anupama Mishra et al. [11] to identify and avert distributed denial of service (DDoS) assaults on cloud servers. To implement our strategy, we relied on feature extraction from statistical data. Based on the findings, it is safe to say that the suggested method can identify DDoS assaults with a low false positive rate and a high accuracy of around 99.68%.
The results obtained by Sukhpreet Singh Dhaliwal et al. [12] demonstrate the efficacy and precision of XGBoost in making dataset predictions. With such a little margin of error, the results are spot on. These results have the potential to be further developed into architecture-ready Intrusion Detection Systems.test data categorization techniques including Support Vector Machines, Random Forest, Neural Networks, Decision Trees, and others. Hybrid models, which combine elements from several approaches, have also recently emerged. After looking at a lot of research, it was found that XGBoost outperforms most categorization algorithms.

Summary:

- XGBoost employs a weighted quantile sketch and an algorithm that takes sparsity into account. Data may be naturally sparse or created by feature engineering artifacts; either way, it manifests as a high number of missing or zero values.
- WQS enables merge and prune operations and is an approximation learning method for trees. LightGBM and XGBoost have shown to be very effective in several machine learning fields.
- When compared to the XGBoost classification approach, Random Forests and hybrid models do reach accuracy that is comparable.

PROPOSED METHOD

The processing power of the cloud resides in a distant data center and makes its services available over an extensive network. All aspects of the program, including data storage, infrastructure, server, and application, are configurable, accessible, and manipulable by the user. Using the internet and the cloud, users may access infrastructure, platforms, and applications from any location in the globe. Communication in the cloud involving two parties. Figure 1 shows that the front end needs to talk to the cloud and users. In order to accomplish tasks such as database maintenance, application development, and service delivery over the network, users may need resources such as hardware and software. Communicating with the cloud and external parties is another need. Virtual computers on the physical layer host the intrusion detection system (IDS). Everything from the site and application servers to the database servers and development tools is taken care of by the third party. There has been a meteoric rise in the use of information technology (IT) in cloud computing across many sectors, including government, businesses, academia, healthcare, and many more. However, they must make sure to carry enough security. Reason being, the cloud is subject to many network attacks. The most common types of conventional assaults are port scanning, user-to-port attacks, distributed denial of service, and IP spoofing [13].
Distributed denial of service (DDoS) attacks were more effective than single-machine assaults because they used a network of infected computers to overwhelm the victim's system. Intruders infiltrate seemingly harmless computers, call them bots or zombies, and give them orders to attack the victim's system.The current state of affairs is plagued by several assaults that deplete existing network infrastructures and compromise security measures including availability, confidentiality, and integrity [11]. Here, we present the RXGB (resistant XGBoost) model, which can identify DDoS attacks (see: figure 1).

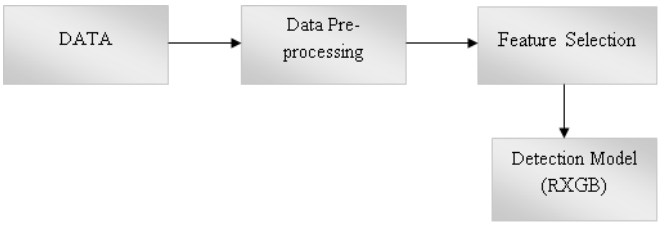


Figure1. Proposed Model Block Diagram

Data

The CICDDOS2019 dataset was acquired by this research from the University of New Brunswick's Canadian Institute of Cybersecurity. For academic study, the institution makes available a comprehensive dataset of DDOS attacks. There are about 1 million legitimate and 30 million malicious traffic flow examples in the collection. There are thirteen distinct types of harmful occurrences. It is 80 features strong and is concerned with the flow of traffic in a network. It was essential to restrict the dataset size used in this investigation due to the enormous size of the CICDDOS2019 dataset [2].There were a total of 32,000 instances in the extracted dataset used for this research; 8,450 of these instances were benign, whereas 23,550 were DDOS assault cases.

Preprocessing

There are two sub-phases in the preprocessing phase: the extraction of packet features and the labeling phase. To build new statistical features, the feature extraction phase captures packets from the network traffic in real-time.For the purpose of detecting and analyzing DDoS attacks, these statistical characteristics are provided in the dataset.It is crucial to identify the network traffic pattern during a distributed denial of service (DDoS) attack and to define the network's quality of service (QoS) based on the discovered attributes. These statistical characteristics are given classes during the labeling process. For each dataset, there are two categories: "1" for attack packets and "0" for nonattack packets. Training the classification tree uses the final dataset, which includes both attack and nonattack data, after labeling.

Feature Selection

After that, we choose features using an ensemble technique. Instance classification performance is improved when different learning algorithms are combined, as shown in ensemble learning, from which ensemble feature selection is developed. An ensemble of feature ranking approaches builds on this previously established idea of an ensemble of learners by integrating many feature ranking methods into a single ranking.

Robust XGBoost (RXGB) for DDOS Anomaly attack detection.

An anomaly based technique is one in which the system is designed to only accept items that exhibit a predetermined pattern of behavior. Any observations that deviate from this pattern are deemed outliers, deviations, or exceptions, and are thus banned from the system. These hidden patterns are uncovered and reported using the anomaly-based detection method [4].

Instances may be classified using decision trees, which are algorithms that use feature values. In a decision tree, each node represents a characteristic of an instance that has to be categorized. The root node is a feature that is chosen to divide the training data. The tree then begins to delve further, creating sub-trees until the same class subsets are generated, using a similar approach. Diagonal portioning is one of the decision trees' constraints. But Zhang proposed additional characteristics using operators like conjunction, negation, and disjunction to build multivariate trees [12].

In the context of ensemble approaches for efficient output prediction, XGBoost is often considered a powerful classifier. XGBoost is a powerful tool for creating guided regression models. With knowledge about the ideal solution (XGBoost) and baseline trainees, however, the validity of this notion may be ascertained. The primary goals in developing XGBoost, which employs the fundamentals of gradient-boosting, were speed and dependability [5].

By using gradient-boosted decision trees, XGBoost primarily aimed to improve performance and speed. It stands for a method for machine boosting, or the application of boosting to machines, which was started by Tianqi Chen and has since been picked up by several developers. The Distributed Machine Learning Community (DMLC) is the proud owner of this product. Tree boosting techniques may make full use of available memory and hardware with the aid of XGBoost, which stands for eXtreme Gradient Boosting. In addition to being used in computer contexts, it improves algorithms and allows for model customization. XGBoost is capable of carrying out all three of the main gradient boosting methods: regularized boosting, stochastic boosting, and gradient boosting. What sets it different from other libraries is its support for adding and adjusting regularization parameters. By making efficient use of memory, the method significantly cuts down on processing time. It has the

unique ability to execute boosting on additional data that is already on the trained model, is Sparse Aware (i.e., it can handle missing values), and enables parallel structure in tree building [12].
The mathematical procedure that yields the predictions yi from the inputs xi is known as the model notion in supervised learning and machine learning more generally. To develop a model that can generate accurate predictions, it is necessary to learn some unknown aspects from data. These factors are called model parameters. To determine these parameters, the model optimization makes use of an objective function, which may be either maximized or reduced.
Both the training loss and the regularization term are essential components of objective functions:

$$\text{obj}(\theta) = L(\theta) + \Omega(\theta) \quad (1)$$

where the regularization term is denoted by **Ω** and L stands for the training loss function. To gauge how well the model can forecast data from the training dataset, we look at the training loss.
We may express the XGboost model mathematically as

$$yi = \sum_{j=1}^{j} fj(xi), fj \in F \quad (2)$$

In this context, j is the total number of trees, and f is a function inside the set of all potential classification and regression trees in the functional space sF [17].
Algorithm1. RXGB Model

LG: The selected learning algorithm
TS: The input data
pit : The packet that was received in real-time from node i
Find node i by using the collection of characteristics xi.
Output: d it : The LG-based node i classification dt
Step 6: Open the incursion dataset. It has 41 characteristics and attributes.
Data pre-processing, which eliminates superfluous and unnecessary information
Selecting features from the virtual machine monitor F1, F2,..., F n
for collection of VM Servers V MSij, i = 1, 2, 3, ...p and j = 1, 2, 3, ...qi do
features SFij of VMSij where SF ij = F1ij, F2ij, F3ij, ...F nij
provide the classifiers the characteristics that were extracted
Apply the RXGB classification model
Create classifier:
Model ← ML Model create(LG, T S).
Generate the predicted decision :
Get Di of node i from the decision history database.
d i t ← predict(Model, pi t)
Return(d i t)
Compute the most frequent decision and select it as the best decision.
Predicted the data, whether the data is normal or abnormal
Disconnect services from that server if action is suspected based on specified rules
otherwise packet is normally forwarded
End

Results

False Positive (FP) denotes alarms that are not true and False Negative (FN) denotes detections that are not true. The properly detected packets are shown by True Positive (TP), while the appropriately rejected packets are shown by True Negative (TN) [16].
The accuracy rate, F-score, and detection rate are the performance measures that will be used to evaluate the proposed IDS. These metrics are derived using the confusion matrix [14].
The proportion of successful predictions relative to the total number of packets in the testing set is called accuracy.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$\text{Detection rate} = \frac{TP}{TP+FN} \quad (4)$$

The F1 score is determined by averaging the accuracy and recall values with a weighting factor. Since this is the case, this score takes into account both false positives and negatives. In situations when there is an unequal distribution of classes, the F1 score is more important, even if precision is harder to compute [15].

$$f-score = \frac{2 \text{ x detection rate x precision}}{\text{Detection rate+ precision}} \quad (5)$$

Confusion matrices include the following elements: true negatives (TN), which mean accurate predictions of normal behavior; true positives (TP), which mean accurate predictions of attack behavior; false positives (FP), which mean incorrect predictions of normal behavior as assault; and false negatives (FN), which mean incorrect predictions of attack as normal.

Table I. Confusion matrix comparison

methodology	**TP**	**FN**	**TN**	**FP**
DT	274	36	245	33
GBT	282	28	253	25
XGB	293	17	265	13
RXGB	302	8	269	9

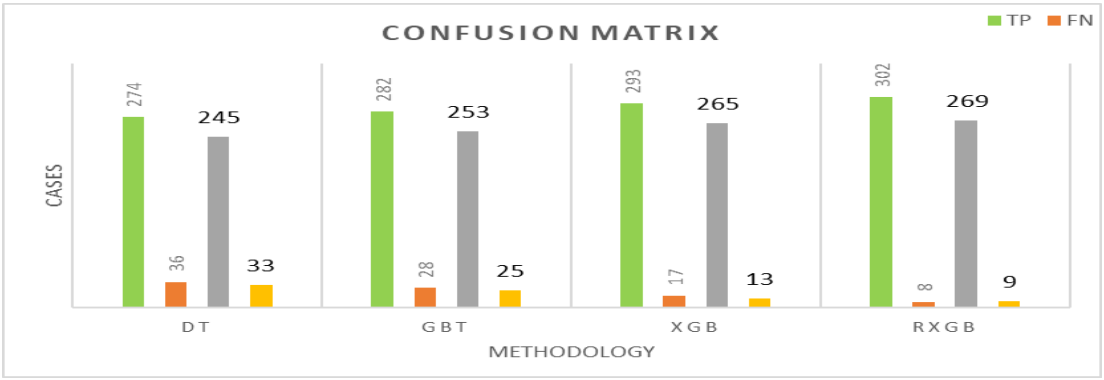


Figure 2. Positive case of Confusion matrix of the proposed and existing method

Table I shows the proposed and current system's confusion parameters, while Figures 2 demonstrate the positive and negative situations, respectively. The method has higher true cases and lower false cases.

Table II. Validation comparison of existing and proposed method.

Methodology	**Precision**	**Recall**	**Fscore**	**Specificity**	**Accuracy**
DT	89.25	88.39	88.82	88.13	88.27
GBT	91.86	90.97	91.41	91.01	90.99
XGB	95.75	94.52	95.13	95.32	94.90
RXGB	97.11	97.42	97.26	96.76	97.11

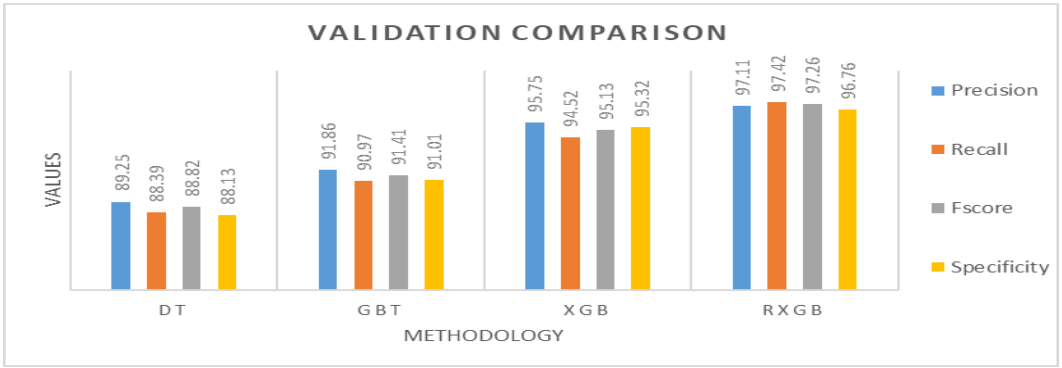


Figure 3. Comparison of validating parameters of proposed and existing system

Table II and Fig. 3 provide the validation parameters (sensitivity, F-score, recall, and precision) for the current and suggested methods. Precision (97.11), recall (97.42), F-score (97.26), and specificity (96.76) are all top-notch metrics that the suggested technique has. The lowest one is seen in DT.

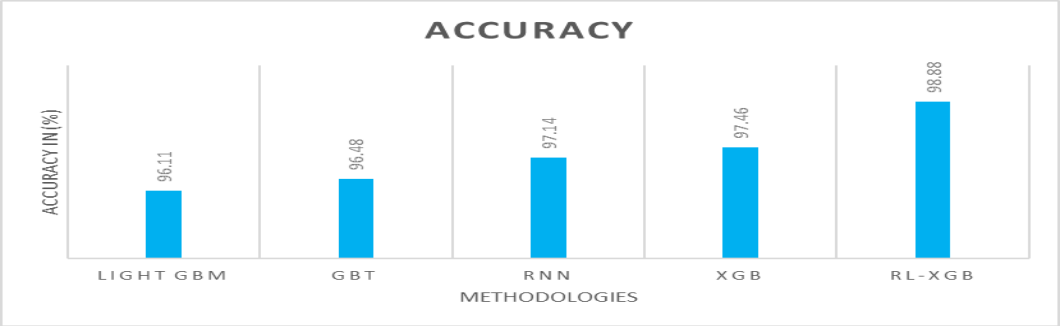


Figure 4. Accuracy comparison

The overall accuracy comparison of the proposed and current approaches Fig. 9. It is evident from the graph above that the suggested strategy outperforms all others and successfully classifies the dataset with a 97.11 % accuracy rate.

CONCLUSION

In addition to failing to identify previously unseen assaults, the increasing sophistication of cyberattacks is rendering traditional Internet security services, like firewalls, useless. In the field of network security, anomaly detection has shown to be crucial for a large business. Fresh threats are appearing on the Internet on a daily basis, necessitating regular updates to detection systems. An innovative RXGB model for anomaly detection is presented in this study. The Robust XGBoost algorithm was used to identify DDoS assaults. According to the findings, our approach is quite effective in detecting DDoS assaults. In our future efforts, we want to apply this technology to a wide range of anomaly detection tasks.

REFERENCES

1. Zekri, M., El Kafhali, S., Aboutabit, N., &Saadi, Y. (2017, October). DDoS attack detection using machine learning techniques in cloud computing environments. In 2017 3rd international conference of cloud computing technologies and applications (CloudTech) (pp. 1-7). IEEE.
2. Alqarni, A. A. (2022). Majority vote-based ensemble approach for distributed denial of service attack detection in cloud computing. Journal of Cyber Security and Mobility, 265-278.
3. Ikram, S. T., Cherukuri, A. K., Poorva, B., Ushasree, P. S., Zhang, Y., Liu, X., & Li, G. (2021). Anomaly detection using XGBoost ensemble of deep neural network models. Cybernetics and information technologies, 21(3), 175-188.
4. Bhati, B. S., Chugh, G., Al-Turjman, F., &Bhati, N. S. (2021). An improved ensemble based intrusion detection technique using XGBoost. Transactions on emerging telecommunications technologies, 32(6), e4076.
5. Raghunath, K. K., Kumar, V. V., Venkatesan, M., Singh, K. K., Mahesh, T. R., & Singh, A. (2022). XGBoost regression classifier (XRC) model for cyber attack detection and classification using inception V4. Journal of Web Engineering, 1295-1322.
6. Naraharisetty, P. P., & Yousef, M. Enhancing Cloud Network Security for IoT Devices: An Integrated Approach with XGBoost and Encryption Techniques.
7. Chen, Z., Jiang, F., Cheng, Y., Gu, X., Liu, W., & Peng, J. (2018, January). XGBoost classifier for DDoS attack detection and analysis in SDN-based cloud. In 2018 IEEE international conference on big data and smart computing (bigcomp) (pp. 251-256). IEEE.
8. Latif, R., Abbas, H., Latif, S., & Masood, A. (2015). EVFDT: an enhanced very fast decision tree algorithm for detecting distributed denial of service attack in cloud-assisted wireless body area network. Mobile Information Systems, 2015.
9. Latif, R., Abbas, H., Assar, S., & Latif, S. (2014). Analyzing feasibility for deploying very fast decision tree for DDoS attack detection in cloud-assisted WBAN. In Intelligent Computing

Theory: 10th International Conference, ICIC 2014, Taiyuan, China, August 3-6, 2014. Proceedings 10 (pp. 507-519). Springer International Publishing.
10. Leevy, J. L., Hancock, J., Zuech, R., &Khoshgoftaar, T. M. (2020, October). Detecting cybersecurity attacks using different network features with lightgbm and xgboost learners. In 2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI) (pp. 190-197). IEEE.
11. Mishra, A., Gupta, B. B., Peraković, D., Peñalvo, F. J. G., & Hsu, C. H. (2021, January). Classification based machine learning for detection of ddos attack in cloud computing. In 2021 ieee international conference on consumer electronics (icce) (pp. 1-4). IEEE.
12. Dhaliwal, S. S., Nahid, A. A., & Abbas, R. (2018). Effective intrusion detection system using XGBoost. Information, 9(7), 149.
13. Nathiya, T., &Suseendran, G. (2018). An effective way of cloud intrusion detection system using decision tree, support vector machine and Naïve bayes algorithm. International Journal of Recent Technology and Engineering, 7, 38-42.
14. Singh, P., &Ranga, V. (2021). Attack and intrusion detection in cloud computing using an ensemble learning approach. International Journal of Information Technology, 13, 565-571.
15. Aldhyani, T. H., &Alkahtani, H. (2022). Artificial Intelligence Algorithm-Based Economic Denial of Sustainability Attack Detection Systems: Cloud Computing Environments. Sensors, 22(13), 4685.
16. Chkirbene, Z., Erbad, A., &Hamila, R. (2019, April). A combined decision for secure cloud computing based on machine learning and past information. In 2019 IEEE Wireless Communications and Networking Conference (WCNC) (pp. 1-6). IEEE.
17. Gouveia, A., &Correia, M. (2020). Network intrusion detection with XGBoost. Recent Advances in Security, Privacy, and Trust for Internet of Things (IoT) and Cyber-Physical Systems (CPS), 137.