International Journal of Advanced Research in Engineering Technology & Science

ISSN: 2349-2819

Email: editor@ijarets.org Volume-9 Issue-1 January-2022 www.ijarets.org

DDOS ATTACK DETECTION: USING MACHINE LEARNING AND THE MUTUAL INFORMATION AND RANDOM FOREST FEATURE IMPORTANCE METHOD

Parula,

Research Scholar, School of Technology and Computer Science,

Glocal University Saharanpur (U. P.)

Dr. Aaruni Goel,

Research Supervisor School of Technology and Computer Science,

Glocal University Saharanpur

ABSTRACT:

Users are provided with on-demand services via the Internet thanks to cloud computing. The services are available at all times and from any location. Despite offering useful services, the paradigm is also vulnerable to security problems. The availability of cloud services is impacted by a Distributed Denial of Service (DDoS) assault, which also poses security risks to cloud computing. DDoS attack detection is essential to maintaining service availability for authorised users. Numerous researchers have investigated the subject, with improved accuracy for various datasets. An approach for DDoS attack detection in cloud computing is presented in this paper. This article's main goal is to make DDoS detection more accurate by reducing misclassification errors. In the proposed study, we utilise two feature selection strategies, namely the Mutual Information (MI) and Random Forest Feature Importance (RFFI) methods, to choose the most pertinent features. Selected features are subjected to Random Forest (RF), Gradient Boosting (GB), Weighted Voting Ensemble (WVE), K Nearest Neighbour (KNN), and Logistic Regression (LR). According to the experimental findings, RF, GB, WVE, and KNN with 19 features have an accuracy of 0.99. Misclassifications of the methodologies are examined for additional investigation, which produces more precise measurements. Numerous tests show that the RF was effective at detecting DDoS attacks and incorrectly identified only one attack as normal. In order to support the suggested method, comparative findings are shown.

Keywords: *machine learning; mutual information; random forest; DDoS; cloud computing*

INTRODUCTION

A platform based on the Internet known as "cloud computing" allows businesses and customers to access servers, databases, and networking on a wide scale while also lowering infrastructure costs [1]. Attackers employ a Distributed Denial of Service (DDoS) assault to bar reputable users from using the services [2]. In this attack, the attackers repeatedly send queries to the victim server, placing a heavy burden on it. The victim server's bandwidth is completely consumed by the attackers' massive amount of requests, rendering it inaccessible to authorised users [3]. The DDoS attack uses a botnet to use brute force on the network's devices, infecting them with malware. Based on the target and behaviour, DDoS attacks can be divided into three primary groups. These include application, traffic, and bandwidth attacks. Attackers use traffic-based attacks to transmit massive amounts of TCP or UDP packets to the victim server, which lowers the victim server's overall performance. In a bandwidth assault, the attackers send a significant amount of anonymous data and increase network traffic. Attackers employ the application assault, which is challenging to mitigate [4,] to target a particular machine. Attack machine learning-based prediction algorithms are used to identify DDoS.

In this paper, we suggest a method for detecting DDoS attacks that makes use of various featureselection and machine learning techniques. We employ the CICIDS 2017 [12] and CICDDoS 2019 [13] datasets, together with the mutual information (MI) and random forest feature importance (RFFI) approaches, to determine which feature is the most pertinent. Attack detection techniques include K-Nearest Neighbour (KNN), Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Weighted Voting Ensemble (WVE). These techniques deliver improved intrusion detection results [14]. Precision, recall, F measure, and accuracy are used to assess the performance of the suggested method. The results demonstrate that the suggested method outperforms existing methods in terms of accuracy with less miss categorization mistakes. By choosing the most pertinent characteristics and performing parameter tweaking on the machine learning model, the main recommendation of this study is to decrease miss classification mistakes in DDoS assault detection.

The following are key contributions.

1. In this study, experiments are performed with tree-based methods (RF, GB), distance based methods (KNN, WVE, and LR), and models based on the CICIDS dataset.
2. This study uses MI and RFFI methods for extraction of the most relevant features.

The rest of the paper is organized as follows.
Section 2 presents different methods for intrusion detection, followed by the proposed methodology in Section 3. Sections 4 and 5 present the results and conclusion, respectively.

MATERIALS AND METHODS

The steps of the suggested methodology for detecting DDoS attacks are detailed in this section. We extract the CICIDS 2017 [12] and CICDDoS 2019 [13] datasets in the first stage. The second phase involves a dataset's preparation. We use machine learning methods to categorise DDoS attacks in the third step. Finally, we assess our method's effectiveness using a variety of measures. The workflow of the suggested methodology for detecting DDoS attacks is shown in Figure 1. The following subsections provide an explanation of each step in the suggested strategy.

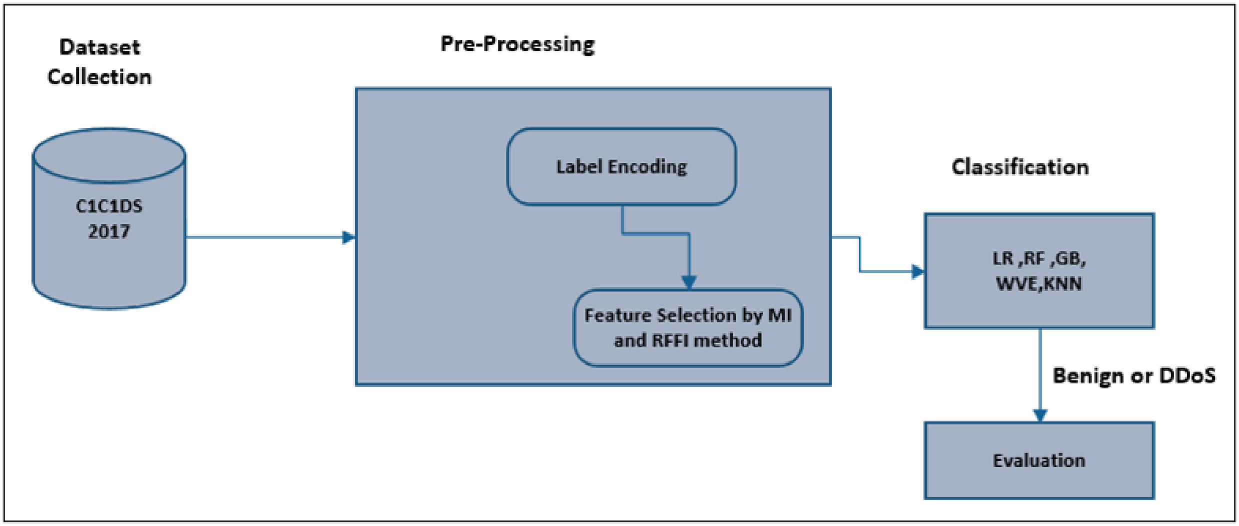

Figure 1. Architecture of the proposed DDoS attack detection model.

Datasets

The datasets for CICIDS 2017 and CICDDoS 2019 were taken from the corresponding websites [12,13]. 3.1 million records of traffic flow make up the CICIDS 2017 dataset [47]. This data set includes traffic flow log files for 5 days, from Monday through Friday. On Friday night's network traffic log file, we performed our experiments. The class label is one of 79 features and 225,711 instances in this log file. The DrDoS_NTP file is chosen as one from the CICDDoS 2019 dataset.

The file cleans 84 input features and has 1,209,961 instances.The benign and DDoS classes make up the binary class label that makes up the class attribute. DDoS is an attack, whereas benign is a regular type. This study's focus is on identifying DDoS assaults from the dataset of other attacks in the other log file. The same data is used for DDoS attack detection in a large number of other studies in the literature. The dataset's abundance of sample data makes it appropriate for assessing detection accuracy.

Data Preprocessing

Preprocessing involves transforming unusable raw data into something that can be used. Apply label encoding to the categorical class label to transform it into discrete form (0,1), where 0 represents a class that is benign and 1 represents a DDoS assault.

1. Feature Selection

The chosen datasets have a high degree of dimension, and as the degree of dimension rises, the training increases exponentially. For the purpose of detecting various attacks, some research have employed feature selection on a chosen dataset [20,25]. A high-dimensional dataset's second drawback is that it makes models more susceptible to over fitting.

The most pertinent features from the features at hand are chosen using a variety of feature selection strategies. The subject of feature selection in data mining and machine learning has received extensive research. An attribute or system that has undergone evaluation is referred to be a feature. Finding the best feature subset of k features that results in the least amount of generalisation error is the aim of feature selection [48]. The filter-based approach, wrapper method, and embedding method are the three primary categories of feature selection. The filter-based method evaluates the link between the input characteristics and the target attribute to determine the relevance of each feature. The wrapper technique creates a model using the subset of features and assesses the model's effectiveness. With millions of instances and huge dimensions, the wrapper method takes longer. The embedded technique chooses the features utilising the knowledge some machine learning models have to provide. A filter-based method of feature selection is MI. When compared to other filter-based techniques, MI has the benefit that it performs well when there is a nonlinear relationship between the input features and the target characteristic. An embedded feature-selection method is the RFFI method. The RFFI method is used in order to produce results that are superior to those produced by other embedded feature-selection techniques. Finding the optimum feature-selection approach for intrusion detection from the filter-based and embedded methods is the goal of employing the MI and RFFI methodologies. The following is a list of the key goals of feature selection**.**

(1) Improve generalization performance, when compared to a model with all characteristics.

(2) Provide more robust generalization and faster reaction to unseen data.

(3) Gain a better and simpler understanding of the data-generation process.

The feature-selection approach is used as a preprocessing step, in regression and classification.

2. Mutual Information

The amount of information that one random variable knows about another random variable is known as MI. Feature selection allows to quantify the importance of a feature subset, in relation to an output vector . Equation (1) shows the calculation of MI.

$$I(X;Y) = H(X) - H(X|Y) \quad \ldots\ldots\ldots..(1)$$

where I(X;Y) is MI for X and Y, H(X) is entropy for X, and H(X|Y) is a conditional entropy for attributes X and Y.

3. Random Forest Feature Importance Method (RFFI)

RF is an ensemble-learning algorithm that grows many decision trees, independently, and combines the output. Decision trees consist of internal and leaf nodes. The selected features are used to make a decision in the internal node, and it divides the dataset into two separate sets, with similar responses. The features in an internal node are selected by the Gini impurity criterion. The feature that has the highest decrease in impurity is selected for the internal node .

DDoS Attack Classification

The following subsections present details of the classification models used. Each model has different parameters that require tuning to achieve better results. This study uses Grid Search (GS) for this purpose.

Logistic Regression

Logistic regression is a machine learning technique that can be used for classification problems. Logistic regression works well on the binary class label. In LR, weights are multiplied with input and pass them to the sigmoid activation function. In the proposed work, we apply LR on selected features for DDoS attack detection. The weights are optimized, using the lbfgs optimizer with C = 0.2.

K Nearest Neighbor

KNN is a classification approach that classifies test data observations, based on how close they are to nearest class neighbors. KNN is used as a semi-supervised learning approach, and KNN is used to identify the nearest neighbors . It is based on a nonparametric approach to classify samples. The distance between separate points on the input vector is determined, and the unlabeled point is, then, allocated to the neighboring class K.

K is the main parameter in the KNN classification. If K is large, the prediction neighbors will take a long time to classify, with an effect on prediction accuracy . KNN is easy to understand, when there are few predictor variables. For the creation of models with normal data types, such as text, KNN is used. We set the value of K as 2, by considering the 2 nearest neighbors, and the Minkowski distance metric is used.

Gradient Boosting

GB is one of the most popular prediction algorithms in machine learning . Various ad hoc parameters are used to regulate the algorithm’s decision tree evolution. Standard regulatory parameters control tree size and weight magnitude. This creates an optimization routine that is free of parameters. However, a variety of parameters are, mostly, used in training, to adjust tree size and shape. Regulation has shown useful results and makes the algorithm constant. Real extreme gradient boosting is a more regularized framework of GB, which has better control regarding the over-fitting issue . As a result, it helps in the prevention of over-fitting in training data. It is linked to a developed set of tools, under the distributed machine learning architecture, due to its efficiency and improved performance. GB has certain parameters that are used in training for DDoS attack detection. Parameters used for GB are shown in Table 1. The parameters are selected on the basis of the GS method used for parameters tuning.

RESULTS AND DISCUSSION

DDoS attack detection and prevention are important problems in a cloud environment. DDoS attack detection is a binary class problem, with benign and DDoS attack class labels. Benign is a normal class. We consider the existence of an attack as a positive class because the interest is in the detection of an attack, and benign is considered as a negative class. MI and RFFI feature selection methods are used. We select 16 features, 19 features, and 23 features, by using the MI and RFFI methods. LR, KNN, GB, RF, and WVE machine learning methods are applied, to selected features. The details of the experimental setup are presented in Table 3. Figures 2–5 show the results of these methods on 16 features,19 features, 23 features, and all features, respectively, on the CICIDS 2017 dataset. The experimental results show that the overall performance of RF is better, compared to other methods in DDoS attacks detection, with 16 features, 19 features, and 23 features. RF, with these features, has a low miss classification rate, compared to other existing methods.

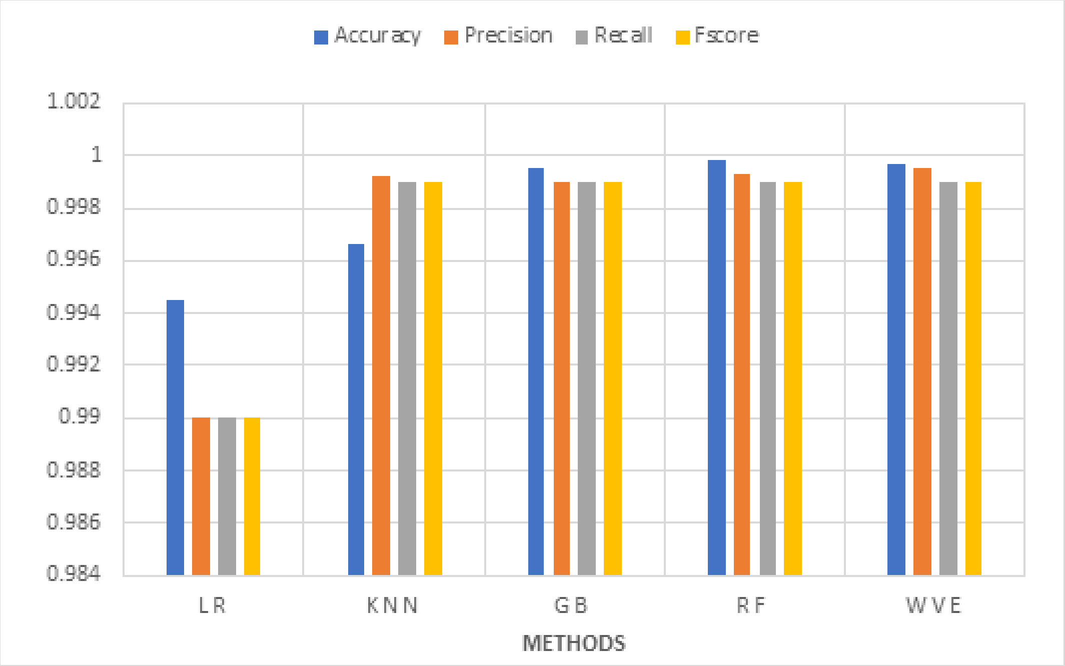

Figure 1. Comparison of different machine learning methods on 16 features.

Figure 1 shows the results of various methods, in DDoS attack detection with 16 features. Sixteen features were selected from the in-hand dataset, using MI and applied machine learning methods, on selected features. RF and WVE methods have the highest prediction accuracy, compared to other methods. All these methods have 99% accuracy and other matrix values. In the case of large datasets, only the measurement of accuracy is not sufficient to measure the performance of the model, since the miss classification of some data points does not affect the accuracy.

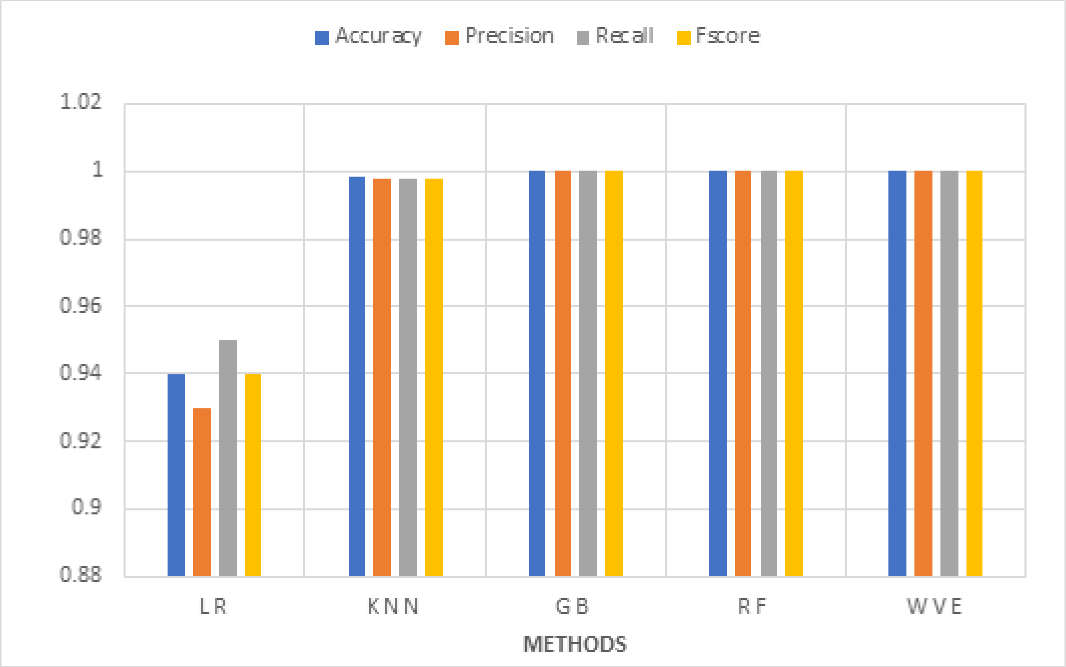

Figure 2. Comparison of different machine learning methods, on 19 features.

Figure 2 shows the results of various machine learning methods in DDoS attack detection, using 19 features that are selected with the RFFI method. The RFFI method is used to select the most relevant feature for DDoS attack detection. The prediction accuracy of the WVE and RF is better, compared to other methods.

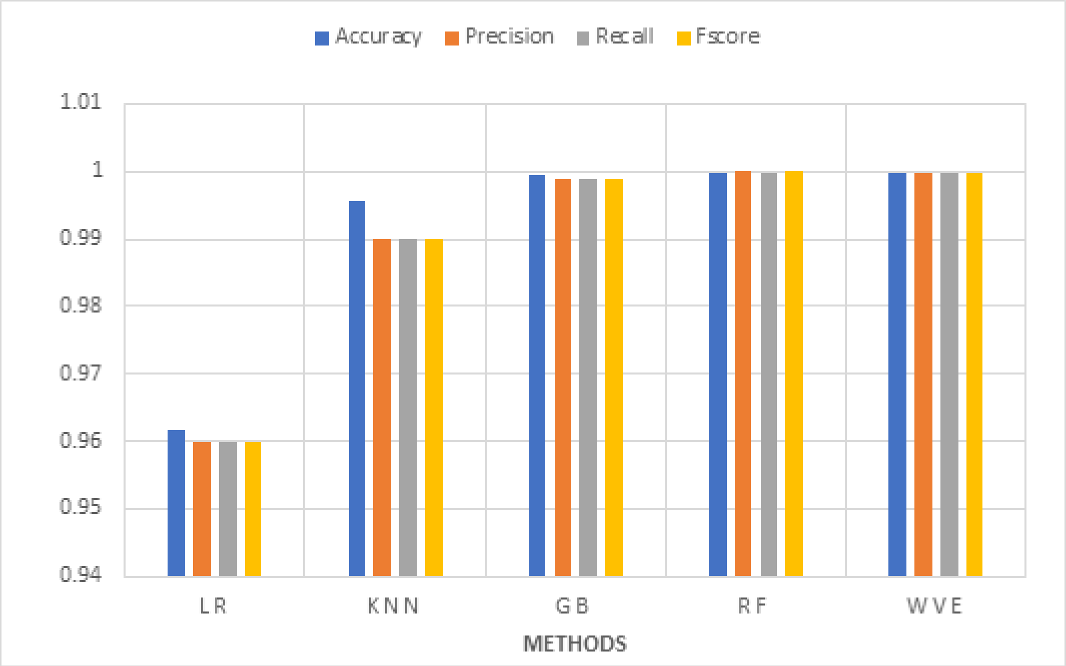

Figure 3. Comparison of different machine learning methods, on 23 features.

Figure 3 shows the results of 23 features, obtained using the MI and applied machine learning methods, on selected features. RF has the highest accuracy, compared to other methods in DDoS attack detection. Twenty-three features were selected from the in-hand dataset, using MI and applied machine learning methods, on selected features. RF has the highest accuracy, compared to other methods in DDoS attack detection.

In comparison to the other approaches used in the detection of a DDoS assault employing 16 features, LR has a high miss classification rate and WVE has a low miss classification rate.

In comparison to GB, KNN, RF, and WVE for DDoS attack classification, LR with 19 features, 23 features, and all features has a high miss classification error. The findings indicate that LR is underperforming at classifying DDoS attacks. On the other hand, the RF and WVE models, which use 19 features, 23 features, and all features, are performing better and have a low miss classification error. These techniques are better appropriate for DDoS attack categorization detection, according to the results.

The comparison between the proposed method's findings and those of the existing methods is shown in Table 6. The proposed method has higher accuracy and a lower miss classification rate than the methods currently in use. The accuracy of the current techniques is close to 99%, although there are more miss categorization errors. With just one such attack, the proposed strategy lowers the miss classification rate. By experimenting with the machine learning approach on various feature sets and by adjusting the machine learning classifier's parameters, the suggested method was able to achieve low miss classification error and high accuracy.

CONCLUSIONS

The detection of DDoS attacks is a frequent issue in a distributed setting. It is crucial to identify this assault since it prevents cloud services from being accessible. An assault of this nature can be recognised using a machine learning model. The goal of this work's research is to more effectively identify DDoS attacks. On the CICIDS 2017 and CICDDoS 2019 datasets, this experiment was run. In trials, several DDoS attack-related files from both datasets were used. By using the MI and RFFI approaches, we choose the traits that are the most pertinent. Machine learning algorithms (RF, GB, WVE, KNN, LR) are fed the chosen features. In comparison to other methods, RF has a greater overall prediction accuracy with 16 features (0.99993) and 19 features (0.999977). It is concluded that employing MI and RFFI as feature selection strategies, RF, GB, WVE, KNN, and LR are producing good results. For the identification of DDoS and other attacks, we may in the future combine wrapper feature selection techniques, such as sequential feature selection, with neural networks.

REFERENCES

1. Malik, N.; Sardaraz, M.; Tahir, M.; Shah, B.; Ali, G.; Moreira, F. Energy-efficient load balancing algorithm for workflow scheduling in cloud data centers using queuing and thresholds. Appl. Sci. **2021**, 11, 5849.
2. Yan, Q.; Yu, F.R. Distributed denial of service attacks in software-defined networking with cloud computing. IEEE Commun. Mag. **2015**, 53, 52–59.
3. Lau, F.; Rubin, S.H.; Smith, M.H.; Trajkovic, L. Distributed denial of service attacks. In Proceedings of the SMC 2000 Conference Proceedings. 2000 IEEE International Conference on Systems, Man and Cybernetics.'Cybernetics Evolving to Systems, Humans, Organizations, and Their Complex Interactions'(Cat. No. 0), Nashville, TN, USA, 8–11 October 2000; IEEE: Piscataway, NJ, USA, Volume 3, pp. 2275–2280.
4. Sambangi, S.; Gondi, L. A Machine Learning Approach for DDoS (Distributed Denial of Service) Attack Detection Using Multiple Linear Regression. Proceedings **2020**, 63, 51.
5. Erickson, B.J.; Korfiatis, P.; Akkus, Z.; Kline, T.L. Machine learning for medical imaging. Radiographics **2017**, 37, 505–515.
6. Hasan, A.; Moin, S.; Karim, A.; Shamshirband, S. Machine learning-based sentiment analysis for twitter accounts. Math. Comput. Appl. **2018**, 23, 11
7. Malik, S.; Tahir, M.; Sardaraz, M.; Alourani, A. A Resource Utilization Prediction Model for Cloud Data Centers Using Evolutionary Algorithms and Machine Learning Techniques. Appl. Sci. **2022**, 12, 2160.
8. Aljamal, I.; Tekeo˘ glu, A.; Bekiroglu, K.; Sengupta, S. Hybrid intrusion detection system using machine learning techniques in cloud computing environments. In Proceedings of the 2019 IEEE 17th International

Conference on Software Engineering Research, Management and Applications (SERA), Honolulu, HI, USA, 29–31 May 2019; IEEE: Piscataway, NJ, USA, 2019, pp. 84–89.

9. Kushwah, G.S.; Ranga, V. Optimized extreme learning machine for detecting DDoS attacks in cloud computing. Comput. Secur. **2021**, 105, 102260.

10. Makuvaza, A.; Jat, D.S.; Gamundani, A.M. Deep Neural Network (DNN) Solution for Real-time Detection of Distributed Denial of Service (DDoS) Attacks in Software Defined Networks (SDNs). SN Comput. Sci. **2021**, 2, 1–10.

11. Manimurugan, S.; Al-Mutairi, S.; Aborokbah, M.M.; Chilamkurti, N.; Ganesan, S.; Patan, R. Effective attack detection in internet of medical things smart environment using a deep belief neural network. IEEE Access **2020**, 8, 77396–77404.

12. Intrusion Detection Evaluation Dataset (CIC-IDS2017). Available online: https://www.unb.ca/cic/datasets/ids-2017.html (accessed on 30 September 2021).

13. DDoS Evaluation Dataset (CIC-DDoS2019). Available online: https://www.unb.ca/cic/datasets/ddos-2019.html (accessed on 27 April 2022).

14. Khan, S.; Kifayat, K.; Kashif Bashir, A.; Gurtov, A.; Hassan, M. Intelligent intrusion detection system in smart grid using computational intelligence and machine learning. Trans. Emerg. Telecommun. Technol. **2021**, 32, e4062.

15. Sandhu, R.S.; Samarati, P. Access control: principle and practice. IEEE Commun. Mag. **1994**, 32, 40–48.

16. Khan, M.S.; Khan, N.M.; Khan, A.; Aadil, F.; Tahir, M.; Sardaraz, M. A low-complexity, energy-efficient data securing model for wireless sensor network based on linearly complex voice encryption mechanism of GSM technology. Int. J. Distrib. Sens. Netw. **2021**, 17, 15501477211018623.

17. Sardaraz, M.; Tahir, M. SCA-NGS: Secure compression algorithm for next generation sequencing data using genetic operators and block sorting. Sci. Prog. **2021**, 104, 00368504211023276.

18. Zhong, Z.; Xu, M.; Rodriguez, M.A.; Xu, C.; Buyya, R. Machine Learning-based Orchestration of Containers: A Taxonomy and Future Directions. ACM Comput. Surv. (CSUR) **2021**.