Implementation of Machine and Deep Learning Algorithms for Intrusion Detection System

Abdulnaser A. Hagar and Bharti W. Gawali

Abstract The intrusion detection system (IDS) is an important aspect of network security. This research article presents an analysis of machine and deep learning algorithms for intrusion detection systems. The study utilizes the CICIDS2017 dataset that consists of 79 features. Multilayer perceptrons (MLPs) and random forests (RFs) algorithms are implemented. Four features extraction techniques (information gain, extra tree, random forest, and correlation) are considered for experimentation. Two models have been presented, the first one using the machine learning random forest (RF) algorithm and the second using deep learning multilayer perceptron (MLP) algorithm. The increased accuracy has been observed when using the random forest algorithm. The RF algorithm gives the best results for the four feature selection techniques, thus proving that RF is better than MLP. The RF algorithm gives 99.90% accuracy, and 0.068% false positive rate (FPR) with 36 features. Furthermore, the dimensionality of the features has been reduced from 79 to 18 features with an accuracy of 99.70% and FRP of 0.19%.

Keywords Machine and deep learning · Random forest and MLP algorithms · Intrusion detection system · Features dimensionality · CICIDS2017 dataset

1 Introduction

Cyber-security is growing to be one among the most significant factors in networks, with the rising progress of computer networks and the huge increasing use of computer applications on these networks. All devices that use networks and the Internet have a threat from the security gaps. The main role of intrusion detection

A. A. Hagar (✉) · B. W. Gawali
Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India
e-mail: csit.hagar@bamu.ac.in

B. W. Gawali
e-mail: bwgawali.csit@bamu.ac.in

G. Rajakumar et al. (eds.), *Intelligent Communication Technologies and Virtual Mobile Networks*, Lecture Notes on Data Engineering and Communications Technologies 131,
https://doi.org/10.1007/978-981-19-1844-5_1

systems is detecting attacks in the networks. Intrusion detection can lead to significant situational attention to online risks, enhance accuracy, and reduce false warnings by linking security events between diverse sources. Investigations have demonstrated that understanding a more different heterogeneous ways of dealing with intrusion detection (ID) improves situational mindfulness and enhances precision. The essential idea of ready connection is that when a similar trademark is causing a similar caution, the system should filter the total numerous cautions into one caution with the goal that a surge of cautions of a similar sort do not happen (rather only a check of those equivalent cautions compose could be accounted for) where alarms are at first related locally in a various leveled form. They are connected again in this manner at a more worldwide level. These connections exercises can include huge preparing power, stockpiling prerequisites, and system activity. Huge variety challenges for ready age can include connection among ready generators, for example IDS that can have a wide range of organizations for their alarm messages or occasion information (usually for associations to have security items with a wide range of restrictive alarm designs, although endeavors are as yet being made to institutionalize). Semantically, cautions can either be viewed as data sources or as yields, as they can likewise fill in as contributions for ready connection purposes. Alarms dependably work at any rate once in a yield limit; however, cautions do not generally work in an info limit [1, 2]. Intrusion detection acts as a crucial part in identifying network attacks. The roles of IDS are the core of protection which gives caution to the system of the network from attacks [3]. A normal intrusion detection system is a freecycle for distinguishing unusual conduct that endeavors to abuse security arrangements in the network. As of late, cybercriminals have demonstrated their place in digital fighting with complex assaults from the attacks. Most of the cyber-security attacks rule the Internet by discouraging the worldwide economy through the burglary of touchy information. Broad exploration has been done in the past to battle cyber-attacks utilizing IDS as the most well-known infrastructure of the network.

Machine learning (ML) and deep learning (DL) are essential tools used to understand how big data is put into intrusion detection system. The enhanced big data requires to be inserted into the intrusion detection system. This can be made possible using machine learning and deep learning that animatedly makes use of certain symmetric machine learning technique. The two forms of machine learning are supervised and unsupervised. The supervised ML has been divided according to the functions as required. Thus, the classification techniques function under the head of supervised machine learning [4, 5]. Furthermore, artificial intelligence (AI) in which ML and DL approaches have additionally been utilized by scientists to identify attacks attributable to their self-learning ability and exactness of forecasts. Deep learning is a sub-area of AI. A multilayer perceptron (MLP) utilizes backpropagation as an administered learning procedure. MLP is a sophisticated learning approach since there are several layers of neurons. As an investigation into computational neuroscience and equally dispersed handling, MLP is frequently used to address difficulties requiring controlled learning. Random forest (RF) is a mainstream AI algorithm that has a place with the supervised learning method. In ML issues, the RF algorithm is utilized for both regression and classification. It depends on the idea of ensemble learning, which is a procedure to combine multiple classifiers toward giving solutions for a problem and to increase the performance of the model [6, 7].

Large volumes of data are handled by an intrusion detection system (IDS) to combat different types of network attacks. Machine learning techniques are employed in this process. However, four strategies (information gain, extra tree random forest, and correlation) are offered to reduce the large dimensionality and features of data, increase accuracy, and lower the false positive rate (FPR) of data analysis. The major challenges faced by big data in intrusion detection are high dimensionality and FPR. The dimensionality of feature reduction states toward techniques, for decreasing the number of features, will be the input for training data. Dimensionality reduction once managing high dimensionality, it is normally valuable to lessen the dimensionality by putting data to a lower-dimensional subspace that catches the 'essence' of the data. Big data is greatly reduced if researchers minimized dimensionality and false positive. When dimensionality and false positive are minimized from intrusion detection big data, researchers can detect various attacks with a faster response and high accuracy. Intrusion detection systems are available in various forms, and there are a variety of techniques to protect your network against attacks. Providing information is necessary after enough data has been gathered.

In the earliest phases of human existence, there were several ways to gather and store information. Hunters communicate the whereabouts of their victims while under their care. An intrusion detection system can deliver improved facts of attacks or intrusion by distinguishing an intruder's actions. Like this way, intrusion detection systems are an influential tool in the administration's fight to keep its calculating resource secure. The basis of the IDS is the parameter of the generation of these categories of intrusion detection systems (IDS). Hybrid, network, and host, work as the basis for which IDS is constructed. There are two more types IDS, such as anomaly depend IDS and signature-based IDS. The environmental network is significantly more important than the performance. The detection of intruders, trespassers, insiders, or man functionaries is made by the hardware as well as a software system in the IDS of the above traditional fire types. Intrusion detection is categorized on a characteristic parameter by the nature of their instructions. These systems are different on the ground of how they detect the intruders and according to the function of their detection. The malfunction may be caused either by misuse or by anomalous use of detection, which is essential to present such measurement. The basics of every intrusion detection system can be more positive or negative. IDS is a useful software to put on each concern. This software supervises and looks after matches closely, cleanly, and shortly for any intrusion interference interposition, breach, and misuse. However, all units are informal of the possible danger. It contains four attack categories [8, 9]:

- Denial of services (DoS): There are various types of attacks involved, e.g., SYN flood. This type of attack is one of the attacks that is prepared by sending a lot of data. DoS attacks build the resources of the host occupied mostly via sending numerous malicious packets that outcomes in the failure of regular network services. It causes a slow pace and results in the DoS. It further causes a device to be out of service. There are numerous types of DOS attacks, for example back, Neptune, pod, smurf, etc. A DDoS attack launched by the attacker comprises mainly of three steps, namely attacking, searching the attack target, and occupying the zombie and actual attacks.

- Remote to Local (R2L): It has unauthorized access from a remote machine, e.g., guessing password. These attacks can occur when the attacker sends packets to a machine over a network but it is not used on that machine. There are several types of R2L attack such as guess_passwd, IMAP, multihop, phf, and ftp_write. An attacker tries to add access to a victim machine without an account, such as a password guessing attack.
- User to Root: It has unauthorized access to local superuser (root) privileges, e.g., various 'buffer overflow' attacks. These types of attacks lead the attacker to start accessing as normal based on the system. An attacker has local access to the victim machine and superuser that attempts to get the privilege. There are some types of user to root attack like, load module, rootkit, and buffer overflow Perl.
- Probing: It is surveillance and probing attack, e.g., port scanning. These attacks take place when the attacker attempts to gather info about the system and network of the computers. An attacker tries to get info about the target host such as ping-sweep, ipsweep, nmap, port sweep, portscan, and Satan.

This research attempts to get an understanding of IDS identification of genuine packets from anonymous packets over the network. Feature selection is likewise identified with reducing the dimensionality of the features which goals to decrease the number of features. Dimensionality reduction is the selection of the most important features. Therefore, this work uses feature selection techniques to reduce dimensionality [10]. A 'false positive' (FP) error occurs when a security system misinterprets a non-malicious activity as an attack. These errors are a critical issue for cyber-security today. Although it might seem that FP errors do not necessarily have serious consequences, incorrect security alerts can lead to significant losses. If unrecognized FP errors occur during training, then the rules which caused them will be incorrectly considered as 'good' and will be used as the foundation for future traffic processing and possibly even future rule development. This can produce cascading error rates. A further complication arises from the relationship between FPs and false negative (FNs) (i.e., attacks that go undetected). When attack-detection thresholds are adjusted to minimize FPs, it tends to increase FNs. Also, the two types of false-alarm errors are asymmetric in their consequences. Generally, FNs incur much higher costs. Therefore, effective FP reduction might increase the overall losses from false alarms [11]. Moreover, the overall objective is addressing the challenges in detecting intrusion which is dimensionality reduction, detecting attacks with high accuracy and less FPR by using ML and DL algorithms.

2 Previous Research Works

IDS is outstanding fields not only for academic research, nonetheless also for cybersecurity research. In recent years, numerous papers have been distributed on this point. In this part, important bits of exploration are discussed. In this section, the related work of the CICIDS2017 dataset, machine learning RF algorithm, and deep learning MLP algorithm are discussed.

Buczak et al. [4] proposed a study of ML approaches that are used by intrusion detection systems. Their work gave three types of classes for the dataset, specifically public datasets, NetFlow data, and packet-level data. Furthermore, it provided a computational complexity for machine learning and mining approaches used through the IDS.

Peng et al. [12] proposed a clustering technique depending on two techniques principal component analysis (PCA) and mini-batch K-means. The PCA technique worked to reduce the dimensionality of features, then the mini-batch K-means ++ technique does the clustering of data. The study used KDDCup1999 to test the work.

Serpil et al. [13] used the random forest for feature reduction, using the CICIDS2017 dataset, by the recursive feature elimination technique. The result of the experiment was accuracy 91% using deep learning MLP. The features reduction was 89% by using the feature elimination technique.

Tang et al. [14] proposed a recurrent neural network for IDS in software-defined network. The authors achieved 89% accuracy in the NSL-KDD dataset. Moreover, for evaluation metrics, accuracy, precision, recall, and F-score were used.

Peng et al. [15] introduced an intrusion detection system depending on THE DECISION TREE algorithm on big data. The authors proposed preprocessing algorithm to detect the string on the KDDCUP99 dataset and used normalization to decrease the input of data to increase the efficiency of their work and improve accuracy. Then, naïve Bayesian algorithm was compared with the decision tree algorithm and KNN algorithm. The result of the decision tree algorithm was found to be the best.

Potluri et al. [16] used machine learning techniques and deep learning techniques to evaluate the performance for detection. The authors used MATLAB and the library of Theano for deep learning. NSL-KDD dataset was used as input data, which have four types of attack (U2R, DoS, R2L, and Probe). The combined softmax regression, SVM, DBN, and stacked autoencoders, called hybrid deep learning, was utilized. After evaluating the proposed hybrid deep learning, the result showed the best accuracy of detection on SVM with stacked autoencoders.

Jiang et al. [17] proposed attack detection on the application layer with the CIC-IDS2017 dataset. It detected the DDoS attack type. The authors implemented two levels, one at the node level and another at the traffic level. At the traffic level, they used features like traffic load, IP counts, average rate, average IP counts, request counts, the average request load, and average order load. The introduced hybrid system that uses deep learning for feature selection increased the accuracy by 99.23%.

Sharafaldin et al. [18] used the RF algorithm for feature selection to determine the family of attack. The work studied the performance for all features with many algorithms which are multilayer perceptron (MLP), AdaBoost, k-nearest neighbor (KNN), quadratic discriminant analysis (QDA), naïve Bayes, Iterative Dichotomiser 3, and random forest (RF). The best precision result obtained with RF and ID3 was 98%.

Potluri et al. [19] used deep learning as a classifier to handle the huge data of network. The dataset used in the work was the NSL-KDD dataset, which had 39 types of attacks that were grouped into four classes of attacks. Their works displayed the result with two classes, one class, normal packet and the another, class attack. The result was 97.7% for accuracy.

Vijayan et al. [20] introduced an IDS and for feature selection, genetic algorithm was used and for classification, and support vector machines were used. Their work was dependent on a linear combination of multiple classifiers of SVM that were sorted according to the attack severity. All classifiers were trained to detect a certain category of attack by using the genetic algorithm for feature selection. They did not use all instances of the CIC-IDS2017 dataset, whereas used few instances.

Watson et al. [15] used deep learning convolutional neural network (CNN) and multilayer perceptron (MLP) algorithm with CIC-IDS2017 dataset. In the study, features were selected from specific packets header features. MLP was the best which produced 94.5% for true positive rate (TPR) and 4.68% for false positive rate (FPR).

3 Dataset

The Canadian Institute of Cyber Security created the dataset of CICIDS2017. The CICIDS2017 dataset is one of the recent datasets for IDS, and it has eight files of CSV that was created in five days. The eight CSV files reflect the attacks that occurred on the five days time (Monday, Tuesday, and Wednesday but Thursday and Friday in the morning and afternoon). Therefore, it includes eight CSV files that show the network traffic profile for each day that contain both normal packets and attack packets. In the tests, CIC-IDS2017 is employed, which is a dataset that meets the eleven essential features of a legitimate IDS dataset: labeling, heterogeneity, metadata, feature set, complete traffic, available protocols, complete network configuration, complete interaction, complete capture, attack diversity, anonymity [18]. CICIDS2017 includes 2,830,743 rows and 79 features. Moreover, it contains a normal packet and fourteen attack types that appear in the Label feature. To make a training and testing subset, the eight files are combined into a single file with a single table including all benign attacks. Then, any features with zeroes as their min and max values, a total of eight features are deleted. Therefore, the features of zero values do not affect any analysis of the dataset. Hence, those features are removed. CIC-IDS2017 contains real-world and the most recent attacks. CICIDS2017 is made by analyzing the traffic of the network utilizing information from the source and destination ports, source and destination IPs, protocols, timestamps, and attacks. Moreover, the CICIDS2017 contains

79 features, but after analysis, 8 features having zeroes values are detected [7]. The dataset is reliable and has the following criteria:

- Includes all protocols.
- Completed network configuration/structure.
- All network traffic is recorded.
- Structure of traffic completed.
- Common attacks are distributed proportionally.

4 Proposed Methodology

By making use of the machine learning RF algorithm and deep learning MLP algorithm with TensorFlow to detect attacks, the efficiency and effectiveness of IDS are increased. Features are selected by four methods, namely information gain, extra tree, random forest, and correlation, and new four datasets are created depending on the number of features for each technique of feature selection. After that, the four datasets enter into two models: one machine learning RF algorithm and another deep learning MLP algorithm. Moreover, the two models evaluate and review the performance matrix. Figure 1 offers the framework of the models.

4.1 Data Preprocessing

Datasets of big data have frequently duplicated features, i.e., noisy, which lead to create challenges for analysis of big data and data modeling, especially in IDS. The CIC-IDS2017 dataset contains eight files, and therefore, after reading all the files using pandas, all files are concatenated to one file. The shape of the dataset becomes 2,830,743 rows, and 79 columns after concatenating. By seeing basic statistical details, it is detected that the min and max values are zeroes for 8 features, which means those features will lead to no effect on any analysis on the dataset, and therefore, those features are removed. After removing those 8 features, the shape of the data set becomes 2,830,743 rows and 71 columns. In addition, the dataset is cleaned from null values [21, 22].

4.2 Feature Selection

When dealing with big data with high dimensionality, the irrelevant and redundant features produce challenges such as decreasing the accuracy of the classification and the effect of the classification process [23, 24]. In big data used for IDS, FS is a preprocessing technique that is widely employed because, in terms of dimensionality, FS is effective [25]. To increase the accuracy and decrease the FPR, the

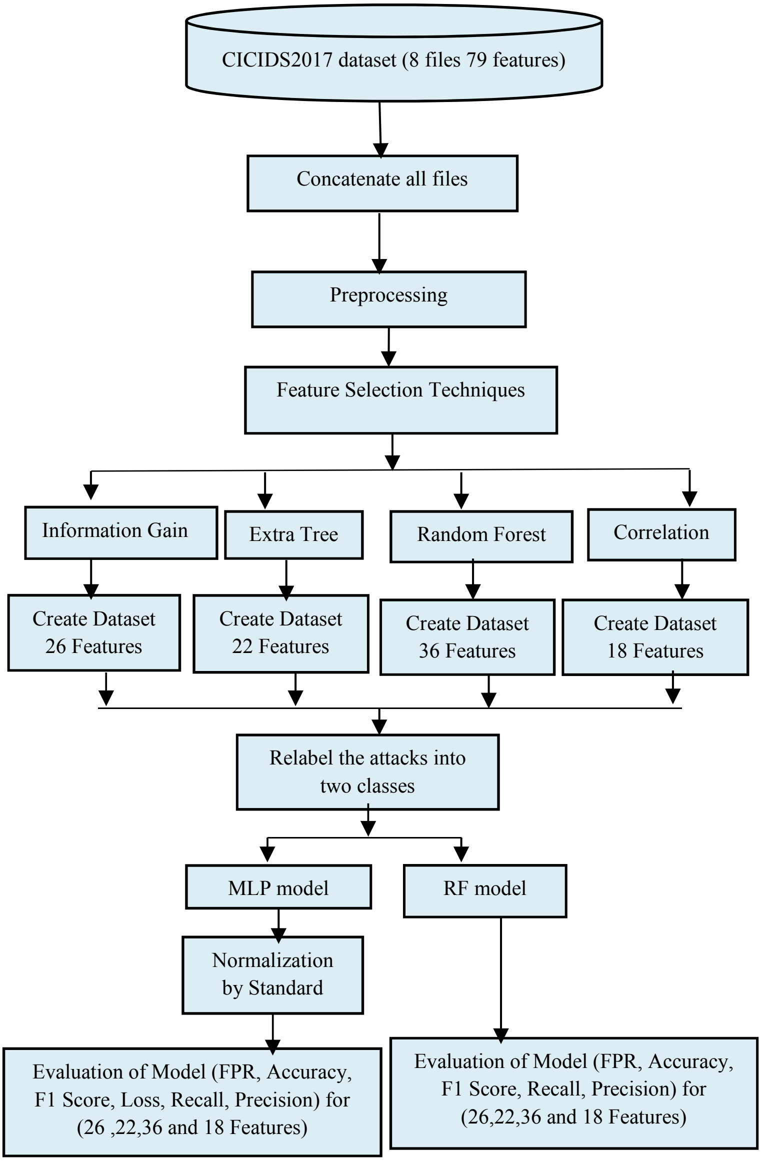
CICIDS2017 dataset (8 files 79 features)
Concatenate all files
Preprocessing
Feature Selection Techniques
Information Gain
Extra Tree
Random Forest
Correlation
Create Dataset 26 Features
Create Dataset 22 Features
Create Dataset 36 Features
Create Dataset 18 Features
Relabel the attacks into two classes
MLP model
RF model
Normalization by Standard
Evaluation of Model (FPR, Accuracy, F1 Score, Loss, Recall, Precision) for (26 ,22,36 and 18 Features)
Evaluation of Model (FPR, Accuracy, F1 Score, Recall, Precision) for (26,22,36 and 18 Features)

Fig. 1 Framework of the proposed work

dimensionality of features is reduced by removing the irrelevant features and redundant features. To identify which feature will be useful in predicting class, FS by four methods (information gain, extra tree, random forest, and correlation) are applied. The features for the technique are:

- Information Gain: It includes the following 26 features; destination port, flow duration, total length (fwd and bwd packets), fwd packet length (max, mean), bwd packet length (max, mean), flow bytes, flow IAT max fwd IAT max, fwd header length, fwd packet, max packet length, packet length (mean, std, variance), average packet size, avg fwd and bwd segment size, fwd header length 1, subflow (fwd, bwd), init win bytes (forward, backward), and label.
- Extra Tree: It includes the following 22 features; destination port, bwd packet length (max, min, mean, std), flow IAT (std, max), fwd IAT (std, max), min packet length, packet length (mean, std, vaiance), push and ACK flag count, average packet size, average bwd segement size, init win bytes forward, min seg size forward, idle (max, min), and label.
- Random Forest: It includes the following 36 features; destination port, total fwd packets, total length (fwd, bwd), fwd packet length (max, mean, std), bwd packet length (max, min, mean, std, flow bytes, flow IAT (std, max), fwd IAT (std, max), fwd and bwd header length, bwd packets, max packet length, packet length (mean, std, variance), psh flag count, average packet size, avg fwd segment size, avg bwd segment size, fwd header length 1, subflow fwd (packets, bytes), subflow bwd bytes, init win bytes (forward, backward, act data pkt fwd, min seg size forward, and label).
- Correlation: It includes the following 18 features; destination port, flow duration, total fwd packets, total length of fwd packets, fwd packet length max, bwd packet length (max, min), flow packets, flow IAT (mean, max), fwd IAT mean, bwd IAT (total, max), fwd psh flags, min packet length, active (std, max), and label.

4.3 Machine Learning

ML is a subset of AI that makes use of statistical learning methods to make sense and make predictions about datasets. In machine learning, there are two types: supervised and unsupervised. A statistical model is trained to predict the output of raw input data using input data and labeled output data in supervised learning. The least-squares approach, for example, generates a straight line (the model) using just a pair of x- and y-values. The line then predicts a y-value (output) for any new x-value (input). An example of supervised learning found in this study uses labeled datasets to predict malicious traffic. The accuracy and performance of different ML algorithms vary depending on the datasets they are applied to. Identifying relevant features and the best method and parameters is the most challenging aspects of employing machine learning in this case [26].

- Random forest (RF) algorithm: The ensemble classifier random forest is used to increase accuracy. A random forest is made up of several different decision trees. When compared to other standard classification methods, random forest has a low classification error. The number of trees, the minimum node size, and the number of characteristics employed to partition each node are all factors to consider [27].

Table 1 Time of execution to select the feature selection and training time

Feature selection techniques	Time of feature selection (min)	Numbers of features	Training time for execution MLP	Training time for execution random forest
Information gain	101.186	26	38.25	21.14
Extra tree	55.35185	22	39.02	13.81
Random forest	387.199	36	39.4019	15.662
Correlation	4.16	18	39.528	14.19

4.4 Deep Learning

Deep learning (DL) was developed in response to advances in technology, research of feature learning, and the availability of enormous amount of labeled data [28].

- Multilayer Perceptron (MLP): The multilayer perceptron is a feed-forward neural network augmentation. It has three layers: an input layer, an output layer, and a concealed layer. MLPs can handle issues that are not linearly separable and are meant to approximate any continuous function. Recognition, pattern classifications, approximation, and prediction are some of MLP's most common applications [29, 30].

4.5 Models

In the proposed models, four FS techniques are implemented. FS by correlation gives the least number of features (18 features) and FS by random forest gives the most features (36 features), while FS by information gain gives 26 features and FS by extra tree gives 22 features as shown in Table 1.

It can be noticed in Table 1 and Fig. 2, the variance in time of selection between the methods, and the correlation method took the least selection time giving the least features, while the random forest took the maximum selection time giving the maximum number of features. Moreover, the extra tree takes less time than information gain and gives the number of features lesser than information gain.

It can be noticed in Table 1 and Fig. 3 that the training time for executing RF has taken around half of MLP training time for executing in the four techniques of feature selection.

Four new datasets are created depending on the number of features, and two models are created; the first model by MLP algorithm (DL) and the second by RF algorithm (ML) as shown in Fig. 1. Moreover, the attacks are relabeled into two classes (normal and attack). The models are carried out by dividing the dataset into train data (66%) and test data (33%) in this study. In each model, four algorithms are used to reduce the dimensionality (26,22,36,18 features) and (2,827,876 samples), which is, 0.66% for training and 0.33% for testing. To achieve this work, 17 programs

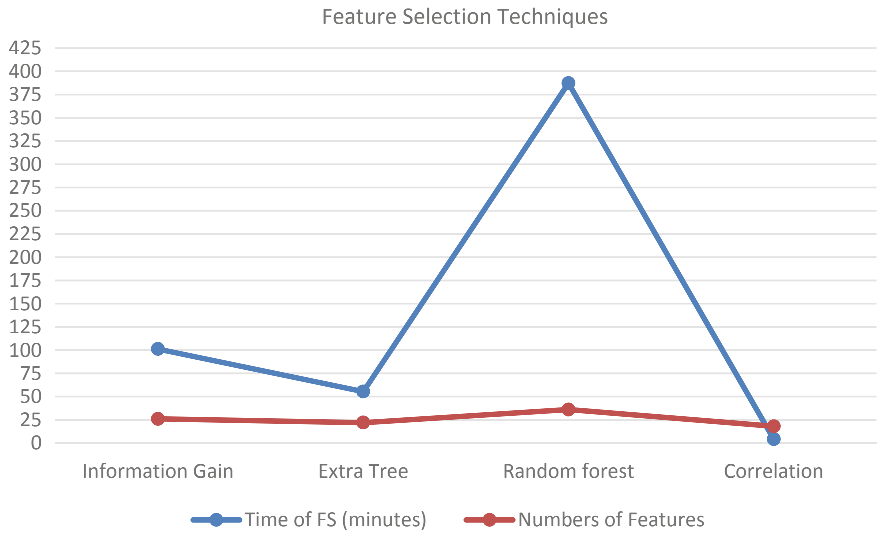

Fig. 2 Time of execution to select FS and numbers of features

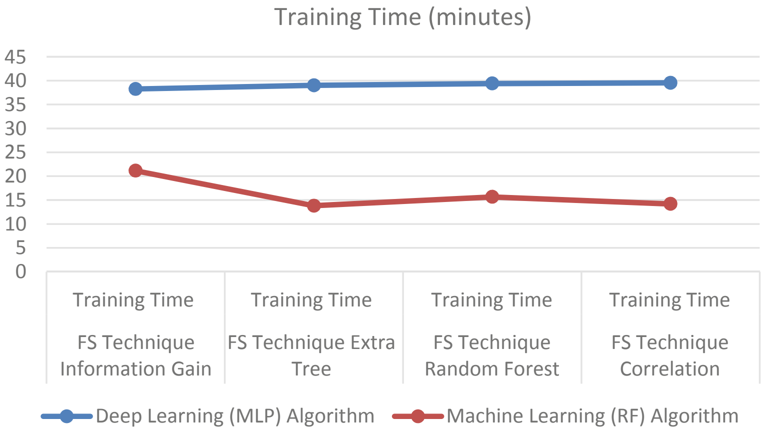

Fig. 3 Training time for execution MLP and random forest

are required to implement (as shown in Table 2) to get the best results for the two models.

Normalization is required on the MLP model. Before appling the algorithm, StandardScaler is chosen for normalization. On the MLP model, hyperparameters set are: (Activation: 'sigmoid,' optimizer: 'rmsprop,', loss function: 'binary_crossentropy,' epochs: 80, and batch size:128).

Table 2 Number of programs for the execution of this work

Number of programs	Purpose of program
1	Concatenating the eight files of CICIDS2017 and preprocessing
4	Feature selection techniques
4	Creating four new datasets depend on the result of feature selection
4	Implementing MLP model for four each new dataset
4	Implementing RF model for four each new dataset
17	Total programs

5 Evaluation of Models

The following are the performance evaluation metrics for the two models [4, 18, 31–34]:

- **Accuracy**: It refers to a model's ability to properly predict both positive and negative outcomes for all predications. It reflects the ratio of the total true negative prediction and true positive prediction from all predictions. The formula to calculate accuracy is TP + TN/(TP + FN + TN + FP).
- **Precision**: The model's precision reflects the model's ability to properly predict positives out of all positive predictions. The chance that a sample labeled as positive is truly positive is measured by precision. The formula to calculate precision is TP/(FP + TP).
- **Recall**: The model's ability to properly forecast positives out of real positives is measured by the model's recall score. The formula to calculate recall is TP/(FN + TP).
- **F1 Score:** The F1 score is the model's score with a function of the precision and recall scores. It may be expressed as the harmonic mean of precision and recall score, the formula to calculate F1 score is, 2*(precision*recall)/(precision + recall).
- **False Positive Rate:** It is the percentage of packets that are accepted as a normal packet but are identified by the system as attack class. The formula to calculate FPR is FP/(FP + TN).
- **False Negative Rate:** It is the percentage of packets identified as an attack nevertheless detected as a normal class by the system. The formula to calculate FNP is FN/(FN + TP).
- **True Positive Rate (Sensitivity):** It is exactly the same as recall, i.e., the percentage of packets with the attack label detected by the system to packets with the same label. The formula to calculate sensitivity is TP/(TP + FN).
- **True Negative Rate (Specificity):** It is the percentage of normal packets label and the packets with the same label that the system has detected. The formula to calculate specificity is TN/(TN + FP).
- **Loss:** Each loss term addresses intra-class variance and inter-class separability together (this extra metric for only deep learning MLP).

TN is the true negative, TP is the true positives, FN is the number of false negatives, and FP is the false positive.

6 Results and Discussion

The result can be determined from Figs. 4, 5, and 6 and Table 3 that the machine learning RF algorithm yield the best result, i.e., using random forest for feature selection technique with 36 features, the results obtained are accuracy 99.90%, precision 99.73%, recall 99.75%, and F1 score 99.74%. The second best results are obtained by RF algorithm and extra tree feature selection technique with 22 features. After that, the RF algorithm and information gain feature selection technique with 26 features. Moreover, the RF algorithm and correlation produced accuracy 99.71%, precision 99.22%, recall 99.30%, and F1 score 99.26% with only 18 features. The features from 79 features are redacted to only 18 features, which led to solving the biggest challenge that is faced by IDS, which is features reduction with high results of all F1 scores, recall, precision, and accuracy. Despite the model of deep learning with MLP algorithm giving results less than the RF algorithm, it still produced high result in the four feature selection techniques. MLP model gave the best result with RF FEATURE SELECTION TECHNIQUE 36 features as shown in Figs. 4, 5, and 6

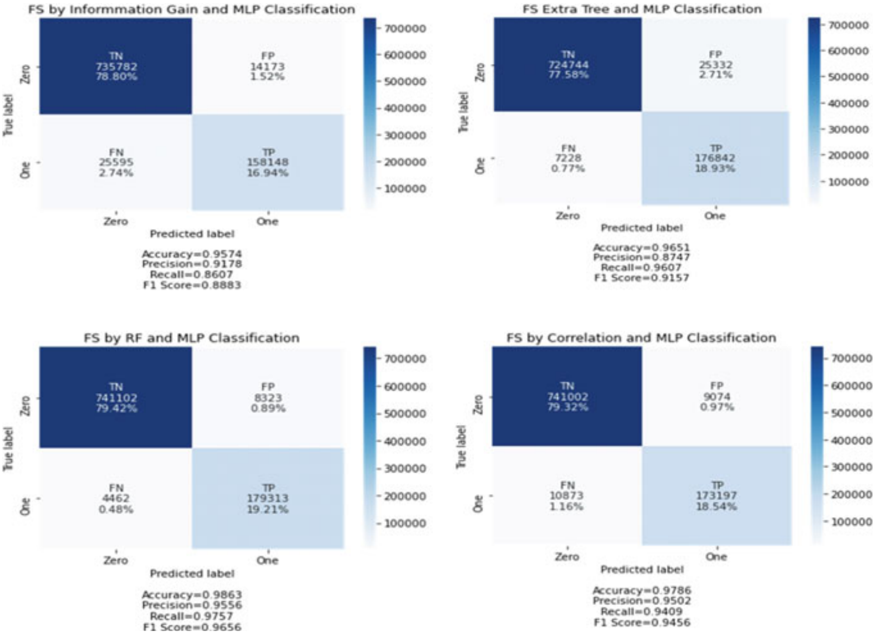

Fig. 4 MLP confusion matrixes

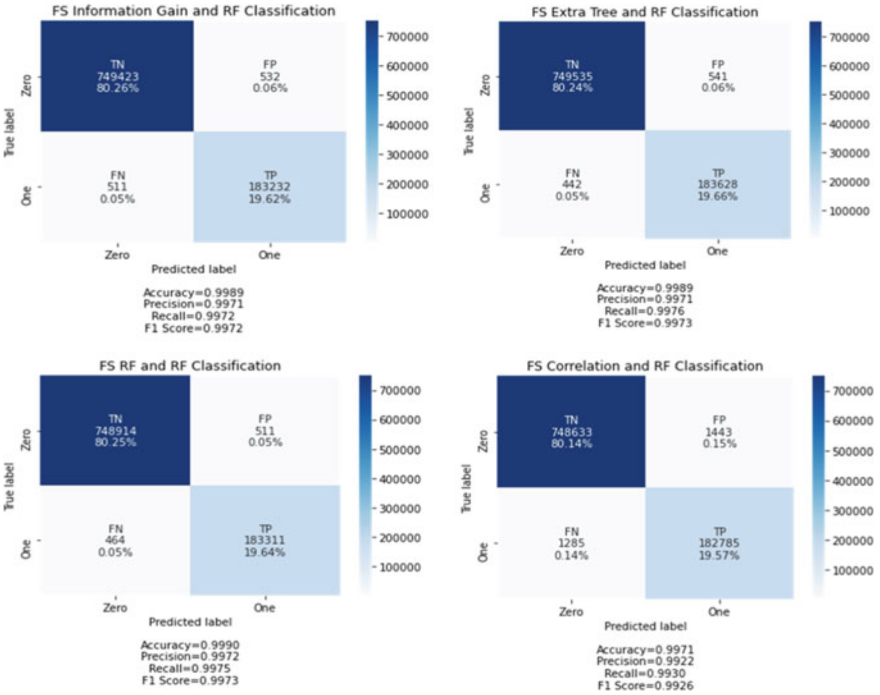

Fig. 5 RF confusion matrixes

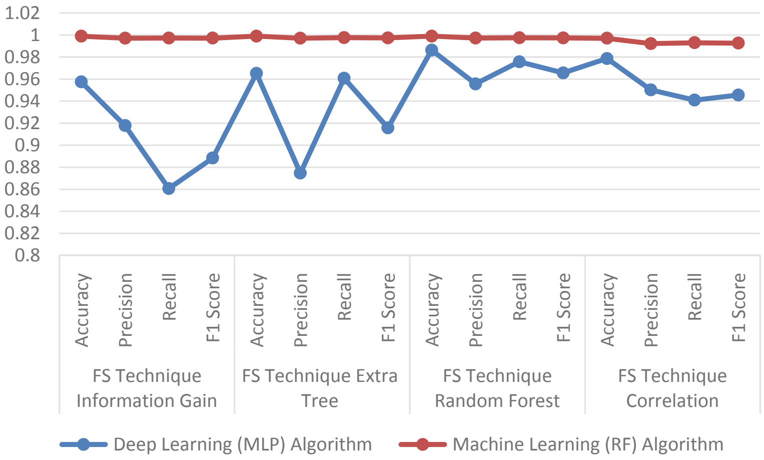

Fig. 6 Evaluation of models

Table 3 Evaluation of models (F1 score, recall, precision, and accuracy)

Feature selection technique	Evaluation metrics	Deep learning (MLP) algorithm	Machine learning (RF) algorithm
Information gain 26 features	Accuracy	0.9574	0.9989
	Precision	0.9178	0.9971
	Recall	0.8607	0.9972
	F1 score	0.8883	0.9972
Extra tree 22 features	Accuracy	0.9651	0.9989
	Precision	0.8747	0.9971
	Recall	0.9607	0.9975
	F1 score	0.9157	0.9973
Random forest 36 features	Accuracy	0.9863	0.9990
	Precision	0.9556	0.9972
	Recall	0.9757	0.9975
	F1 score	0.9656	0.9973
Correlation 18 features	Accuracy	0.9786	0.9971
	Precision	0.9502	0.9922
	Recall	0.9409	0.9930
	F1 score	0.9456	0.9926

and Table 3, i.e., accuracy 98.63%, precision 96.56%, recall 97.57%, and F1 score 96.56%. After evaluating the result, one of the most challenges IDS face, which is features reduction by reduction of features from 79 to 36, 22, 26, and 18 feature with high results of all evaluation metrics for models (F1 score, recall, precision, and accuracy) has been addressed.

As shown in Figs. 7, 8 and Table 4, it can be noticed that the result of FPR is 0.068% by RF model with random forest for feature selection technique (36 features), while the MLP model gave FPR 1.11%. Furthermore, the RF result of FNR gave 0.25% and MLP gave FNR 2.4%. After evaluating the results, it is evident that one of the most challenges that IDS face, which are FPR and FNR, has been addressed.

From Table 5 and Figs. 9 and 10, it is clear that the best result is obtained by RF model with random forest feature selection, whose sensitivity is 99.75% and specificity 99.93%.

From Table 6 and Figs. 11 and 12, it is noticed that the MLP model with feature selection random forest gave the best results as accuracy 98.63%, and loss 5.51% from the four feature selection techniques.

From all the above results, it is noticed that the machine learning random forest algorithm gave results better than the deep learning MLP algorithm due to the data set labeled attack (supervised) and because the random forest is a predictive modeling tool rather than a descriptive one. However, alternative techniques would be appropriate if a description of the relationships in data is required.

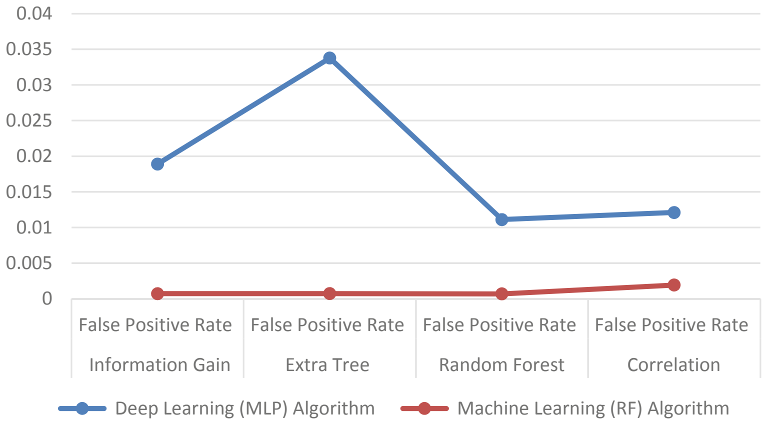

Fig. 7 False positive rate

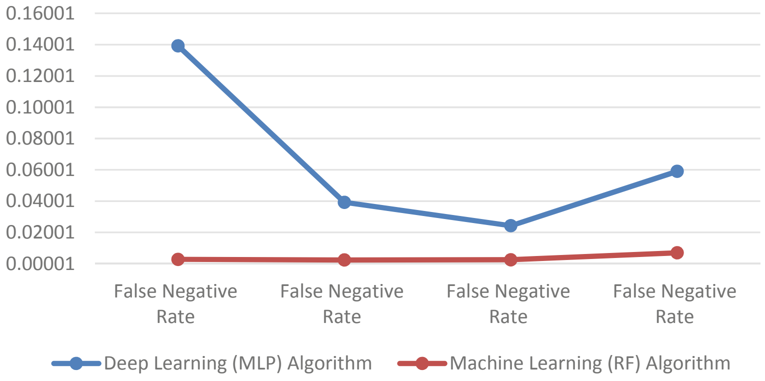

Fig. 8 False negative rate

Table 4 False positive rate and false negative rate

Feature selection technique	Evaluation metrics (FPR, and FNR)	Deep learning (MLP) algorithm	Machine learning (RF) algorithm
Information gain 26 features	FPR	0.0189	0.00071
	FNR	0.1393	0.00279
Extra tree 22 features	FPR	0.0338	0.00072
	FNR	0.0393	0.00240
Random forest 36 features	FPR	0.0111	0.00068
	FNR	0.0243	0.00252
Correlation 18 features	FPR	0.0121	0.00192
	FNR	0.0591	0.00698

Table 5 Sensitivity (TPR) and specificity (TNR))

Feature selection	Evaluation metrics	Deep learning (MLP) algorithm	Machine learning (RF) algorithm
Information gain 26 features	Sensitivity	0.8607	0.9972
	Specificity	0.9811	0.9993
Extra tree 22 features	Sensitivity	0.9607	0.9976
	Specificity	0.9662	0.9993
Random forest 36 features	Sensitivity	0.9757	0.9975
	Specificity	0.9889	0.9993
Correlation 18 features	Sensitivity	0.9409	0.9930
	Specificity	0.9879	0.9981

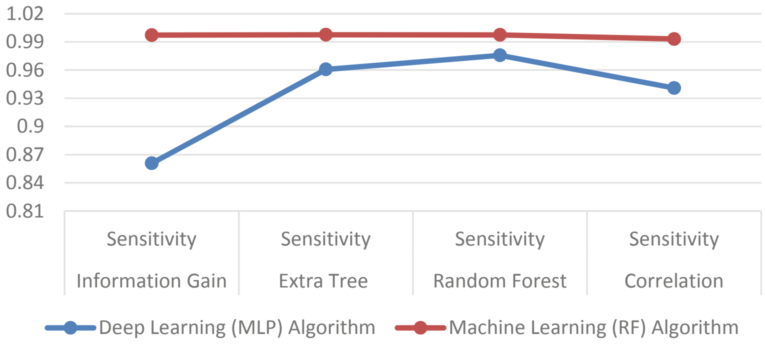

Fig. 9 Sensitivity (TPR)

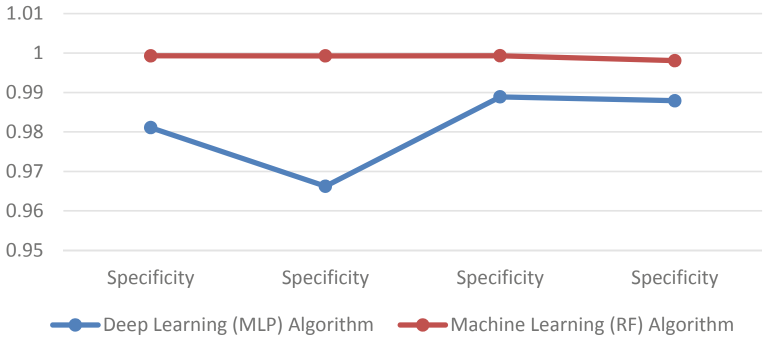

Fig. 10 Specificity (TNR)

Table 6 Feature selection techniques and classification by MLP

Feature selection techniques and results by MLP model

Evaluation metric for MLP model	FS technique ınformation gain	FS technique extra tree	FS technique random forest	FS technique correlation
Accuracy	0.9574	0.9651	0.9863	0.9786
Loss	0.1045	0.0822	0.0551	0.0534

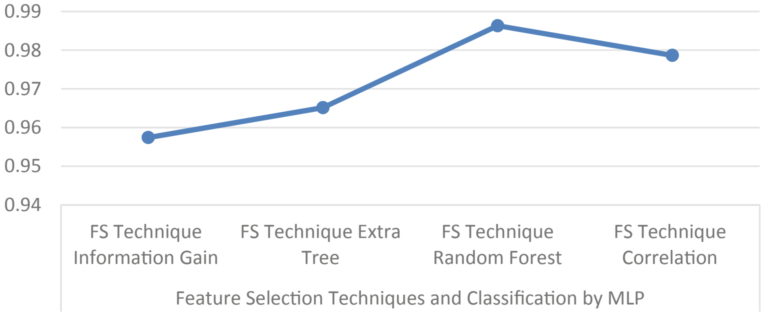

Fig. 11 MLP accuracy

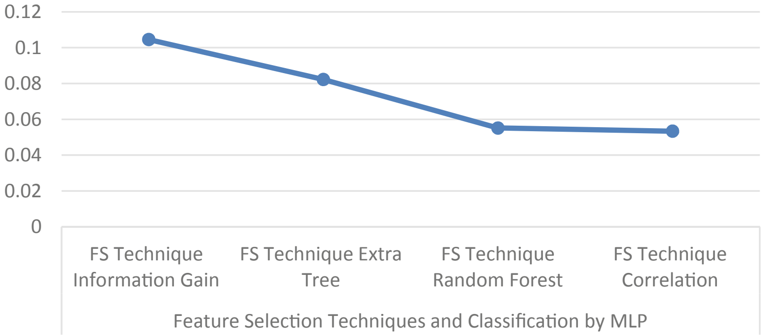

Fig. 12 MLP losses

7 Conclusion

This research work presents the utilization of ML and DL algorithms on the CICIDS2017 dataset. The work is performed on four feature selection techniques (information gain, extra tree, random forest, and correlation). For evaluation and classification of normal and attacked packets, two models, i.e., deep learning model and machine learning model, have been proposed. The accuracy has been increased, and the FPR has been decreased by using the deep learning MLP algorithm and machine learning RF algorithm. RF algorithm gave the best result of accuracy 99.90% and FPR

0.068%, while MLP gave accuracy 98.63% and FPR 1.11%. Moreover, the dimensionality of the dataset is reduced from 79 to 18 features with 99.70% accuracy and 0.19% FPR.

References

1. Vinayakumar R, Alazab M, Soman KP, Poornachandran P, Al-Nemrat A, Venkatraman S (2019) Deep learning approach for ıntelligent ıntrusion detection system. IEEE Access 7:41525–41550. https://doi.org/10.1109/ACCESS.2019.2895334
2. Abdulraheem MH, Ibraheem NB (2019) A detailed analysis of new intrusion detection dataset. J Theor Appl Inf Technol 97(17):4519–4537
3. Hagar AA, Chaudhary DG, Al-bakhrani ALIA, Gawali BW (2020) Big Data analytic using machine learning algorithms for intrusion detection system: a survey, vol 10, no 3, pp 6063–6084
4. Buczak AL, Guven E (2016) A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Commun Surv Tutorials 18(2):1153–1176. https://doi.org/10.1109/COMST.2015.2494502
5. Sathesh A (2019) Enhanced soft computıng approaches for intrusion. J Soft Comput Paradigm 1(2):69–79
6. Farhan RI, Maolood AT, Hassan NF (2020) Performance analysis of flow-based attacks detection on CSE-CIC-IDS2018 dataset using deep learning. Indonesian J Electr Eng Comput Sci 20(3):1413–1418. https://doi.org/10.11591/ijeecs.v20.i3.pp1413-1418
7. Karatas G, Demir O, Sahingoz OK (2020) Increasing the performance of machine learning-based IDSs on an imbalanced and up-to-date dataset. IEEE Access 8:32150–32162. https://doi.org/10.1109/ACCESS.2020.2973219
8. Joe CV, Raj JS (2021) Deniable authentication encryption for privacy protection using blockchain. J Artif Intell Capsule Netw 3(3):259–271
9. Goeschel K (2016) Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and Naive Bayes for off-line analysis. In: Conference Proceedings—IEEE SOUTHEASTCON, vol 2016. https://doi.org/10.1109/SECON.2016.7506774
10. Leevy JL, Khoshgoftaar TM (2020) A survey and analysis of intrusion detection models based on CSE-CIC-IDS2018 Big Data. J Big Data 7(1). https://doi.org/10.1186/s40537-020-00382-x
11. Almansob SMH, Lomte SS (2017) Addressing challenges in big data intrusion detection system using machine learning techniques. Int J Comput Sci Eng 5(11):127–130. https://doi.org/10.26438/ijcse/v5i11.127130
12. Peng K, Leung VCM, Huang Q (2018) Clustering approach based on mini batch Kmeans for ıntrusion detection system over Big Data. IEEE Access 6:11897–11906. https://doi.org/10.1109/ACCESS.2018.2810267
13. Ustebay S, Turgut Z, Aydin MA (2018) Intrusion detection system with recursive feature elimination by using random forest and deep learning classifier. In: 2018 International congress on big data, deep learning and fighting cyber terrorism, pp 71–76
14. Tang TA, Ali S, Zaidi R, Mclernon D, Mhamdi L, Ghogho M (2018) Deep recurrent neural network for ıntrusion detection in SDN-based networks
15. Peng K, Leung VCM, Zheng L, Wang S, Huang C, Lin T (2018) Intrusion detection system based on decision tree over big data in fog environment. Wirel Commun Mob Comput 2018. https://doi.org/10.1155/2018/4680867
16. Potluri S, Henry NF, Diedrich C (2017) Evaluation of hybrid deep learning techniques for ensuring security in networked control systems

17. Jiang J et al (2018) IEEE International conference on big data science and engineering method for application layer DdoS. In: 2018 17th IEEE International conference on trustworthy security and privacy computer communication. 12th IEEE International conference on big data science and engineering, pp 1565–1569 (2018). https://doi.org/10.1109/TrustCom/BigDataSE.2018.00225
18. Sharafaldin I, Lashkari AH, Ghorbani AA (2018) Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: ICISSP 2018—Proceedings of 4th International conference on ınformation systems, security and privacy, vol 2018, no Cic, pp 108–116. https://doi.org/10.5220/0006639801080116
19. Potluri S, Diedrich C (2016) Accelerated deep neural networks for enhanced ıntrusion detection system
20. Vijayanand R, Devaraj D, Kannapiran B (2018) Intrusion detection system for wireless mesh network using multiple support vector machine classifiers with genetic-algorithm-based feature selection. Comput Secur. https://doi.org/10.1016/j.cose.2018.04.010
21. Stiawan D, Yazid M, Bamhdi AM (2020) CICIDS-2017 dataset feature analysis with ınformation gain for anomaly detection. IEEE Access XX:1–12. https://doi.org/10.1109/ACCESS.2020.3009843
22. Abdulhamed R et al (2019) Features dimensionality reduction approaches for machine learning based network. Electronics. https://doi.org/10.3390/electronics8030322
23. Hamid Y, Balasaraswathi VR, Journaux L, Sugumaran M (2018) Benchmark datasets for network intrusion detection: a review. Int J Netw Secur 20(4):7. https://doi.org/10.6633/IJNS.2018xx.20(x).xx
24. Othman SM, Ba-Alwi FM, Alsohybe NT, Al-Hashida AY (2018) Intrusion detection model using machine learning algorithm on Big Data environment. J Big Data 5(1). https://doi.org/10.1186/s40537-018-0145-4
25. Keerthi Vasan K, Surendiran B (2016) Dimensionality reduction using principal component analysis for network intrusion detection. Perspect Sci 8:510–512. https://doi.org/10.1016/j.pisc.2016.05.010
26. Zhou L, Pan S, Wang J, Vasilakos AV (2017) Machine learning on big data: opportunities and challenges. Neurocomputing 237:350–361. https://doi.org/10.1016/j.neucom.2017.01.026
27. Genuer R, Poggi JM, Tuleau-Malot C, Villa-Vialaneix N (2017) Random forests for big data. Big Data Res 9:28–46. https://doi.org/10.1016/j.bdr.2017.07.003
28. Chockwanich N, Visoottiviseth V (2019) Intrusion detection by deep learning with tensorflow. In: International conference on advanced communication technology (ICACT), vol 2019, pp 654–659. https://doi.org/10.23919/ICACT.2019.8701969
29. Abirami S, Chitra P (2020) Energy-efficient edge based real-time healthcare support system, 1st edn, vol 117, no 1. Elsevier
30. Basnet RB, Shash R, Johnson C, Walgren L, Doleck T (2019) Towards detecting and classifying network intrusion traffic using deep learning frameworks. J Internet Serv Inf Secur 9(4):1–17. https://doi.org/10.22667/JISIS.2019.11.30.001
31. Wang L, Jones R (2017) Big data analytics for network intrusion detection: a survey. Int J Netw Commun 7(1):24–31. https://doi.org/10.5923/j.ijnc.20170701.03
32. Dahiya P, Srivastava DK (2020) Intrusion detection system on big data using deep learning techniques. Int J Innov Technol Exploring Eng 9(4):3242–3247. https://doi.org/10.35940/ijitee.D2011.029420
33. Fernandes G, Carvalho LF, Rodrigues JJPC, Proença ML (2016) Network anomaly detection using IP flows with principal component analysis and ant colony optimization. J Netw Comput Appl 64:1–11. https://doi.org/10.1016/j.jnca.2015.11.024
34. Kato K, Klyuev V (2017) Development of a network intrusion detection system using Apache Hadoop and Spark. In: 2017 IEEE conference on dependable security and computing, pp 416–423. https://doi.org/10.1109/DESEC.2017.8073860