International Journal of Multidisciplinary Research and Growth Evaluation.

Intrusion detection system using machine learning algorithm

Jaya Varshini R [1*], Sifa Thahasin F [2], Jayasri S [3], Kannan N [4]
[1-3] Computer Science and Engineering, EGS Pillay Engineering College, Nagapattinam, Tamil Nadu, India
[4] Assistant Professor, Computer Science and Engineering, EGS Pillay Engineering College, Nagapattinam, Tamil Nadu, India

* Corresponding Author: **Jaya Varshini R**

Article Info

ISSN (online): 2582-7138
Volume: 04
Issue: 03
May-June 2023
Received: 25-03-2023;
Accepted: 18-04-2023
Page No: 31-36

Abstract
Intrusion Detection System (IDS) is meant to be a software application which monitors the network or system activities and finds out any malicious operation occurs. Tremendous growth and usage of internet raises concerns about how to protect and communicate the digital information in a safe manner. Nowadays, hackers use different types of attacks for getting the valuable information. As the internet emerging into the society, new stuffs like viruses and worms are imported. The malignant, the users use different techniques like cracking of password, detecting unencrypted text are used to cause vulnerabilities to the system. Hence, security is needed for the users to secure their system from the intruders. Firewall technique is one of the popular protection techniques and it is used to protect the private network from the public network. IDS is used in network related activities, medical applications, credit card frauds, Insurance agency. Many intrusion detection techniques, methods and algorithms help to detect those attacks. The main objective of this project is to provide a comparative study about intrusion detection using various machine learning and deep learning techniques. Various machine learning techniques have been used to develop IDs, such as Random Forest algorithm, Support Vector Machine and Gradient boosting algorithm in real time network datasets such as Intrusion Detection System (IDS) datasets and UNSW datasets. Gradient boosting is a popular machine learning algorithm that can be used for intrusion detection in computer networks. The algorithm involves iteratively adding weak learners to a model to improve its overall predictive power. The proposed system can be analyzed in terms of error rate and accuracy values and implement in python tool for performance analysis.

Keywords: Intrusion Detection System, Machine Learning, Attack Detection, Performance Analysis

1. Introduction
An Intrusion Detection System (IDS) [1], a security technology designed to detect and alert on unauthorized access or attacks on a computer network. Large network traffic sizes, highly unequal data transmission, the difficulty of distinguish between natural and irregular operation, and the need of constant adaption to an even changing environment are all obstacles that an IDS (Intrusion Detection System) [1]. In general, the challenge is to effectively identify and classify various operations in a computer network. The most two general form of intrusion detection methods used to search for detection. A Software Defined Network (SDN) has characteristics such as programmability, centralized control, and global view, so it is widely used in Network Security. Previous studies [2, 3, 4, 5] proposed the use of SDN technology to redirect network traffic to snort IDS for detecting malicious attacks.
Snort, a signature-based detection system, cannot detect unknown attacks and adapt to large-scale traffic. On the contrary, anomaly detection, developed as classifiers to differentiate anomalous traffic from normal traffic, is well suited for the detection of unknown attacks, but it has a high false alarm rate. Misuse recognition techniques use signature matching algorithm [6] to search for known instances of misuse of both network and system behaviour.

Many types of shallow discriminative machine learning techniques have been extensively applied to IDS, such as Neural Network (NN), Random Forest (RF), Decision Tree (DT), and Support Vector Machine (SVM).

An IDS approach is effective for detecting previously known threats. The IDS can generate warning, but reacting to each one consumes time and energy allowing the system to become unreliable. IDS should not begin the removal process as soon as the first symptom is detected, but should instead wait until all of the warnings have been received before making a division based on their data. However, these approaches provide unsatisfactory classification or detection accuracy. The intrusion detection result depends not only on the performance of the classifier but also on the quality of the input data. Network traffic data usually involves high dimensionality and redundancy of features, which can easily cause a feature dimensionality disaster. Therefore, feature dimensionality reduction is particularly important for effectively improving the performance of the above mentioned supervised classifiers [7]. It includes two types of techniques: feature subset selection and feature extraction [8]. Feature subset selection works by removing relevant or redundant features; the subset of features selected will give the best performance according to some objective function.

Many studies [9, 10, 11, 12, 13] have demonstrated that feature selection methods can overcome the "dimensionality curse" and achieve high detection performance in CIDS. Feature extraction maps the original high dimensional features into low dimensional features and generates new linear or non-linear combinations of the original features [8]. Recently, various researchers have demonstrated that deep learning technology has considerable potential for IDS, especially in feature extraction.

Aims to use the machine learning technique to automatically extract essential features from raw network data and input them into a shallow classifier for effective identification of attacks. Anomaly detection systems [14] concentrate on the development of a normalized model of user interaction. Anomaly detection strategy is more effective at identifying novel threats an intruder. A network IDS scans incoming network traffic for patterns that suggest for infect computer data on the computer.

2. Related Works

Shadi Aljawarneh, Monther Aldwairi, Muneer Bani Yassein *et al* [15] proposed hybrid model for dimensionality reduction improves the accuracy rate and reduces the detection time. The analysis performed on the NSL-KDD dataset through the help of tables and figures has allowed the researcher to gain a clearer dataset understanding. However, there are issues with obtaining high false and low false negative rates. A hybrid approach with two main parts is proposed to address these issues. First, data needs to be filtered using the Vote algorithm with Information Gain that combines the probability distributions of these base learners in order to select the important features that positively affect the accuracy of the proposed model. Next, the hybrid algorithm consists of following classifiers: J48, Meta Pagging, Random Tree, REP Tree, AdaBoostM1, Decision Stump and Naïve Bayes. Based on the results obtained using the proposed model, observed a improved accuracy, high false negative rate, and low false positive rule.

Nasrin Sultana, Naveen Chilamkurti and Rabei Alhadad *et al* [16] provided an overview of programmable networks and examined the emerging field of Software Defined Networking (SDN). Outlined various intrusion detections mechanisms using ML/DL approaches. Emphasized software defined networking (SDN) technology as a platform using ML/DL approaches to detect vulnerabilities and monitor networks. Software Defined Networking Technology (SDN) provides a prospect to effectively detect and monitor network security problems ascribing to the emergence of the programmable features. Machine Learning (ML) approaches have been implemented in the SDN-based Network Intrusion Detection Systems (NIDS) to protect computer networks and to overcome network security issues. A stream of advanced machine learning approaches the deep learning technology (DL) commences to emerge in the SDN context.

Peng, Leung and Huang *et al* [17] proposed a clustering method based on Mini Batch Kmeans with PCA (PMBKM) for Intrusion Detection System. Taking IDS classic dataset KDDCUP99 for example, both 10% dataset and full dataset are tested. Firstly, we pre-process the given dataset and then the PCA method is used to reduce the dimension so as to improve the clustering efficiency. Additionally, the Mini Batch Kmeans method is used for the clustering of the processed dataset. Compared with Kmeans (KM), Kmeans with PCA (PKM), as well as Mini Batch Kmeans (MBKM), the experimental results show that our proposed PMBKM is effective and efficient. Above all, PMBKM can be used for intrusion detection system over big data environment. In our future work, we will engage in the research of clustering method over fog computing. First, a pre-processing method is proposed to digitize the strings and then the data set is normalized so as to improve the clustering efficiency. Second, the principal component analysis method is used to reduce the dimension of the processed data set aiming to further improve the clustering efficiency, and then mini batch K-means method is used for data clustering. More specifically, used K-means++ to initialize the centers of cluster in order to avoid the algorithm getting into the local optimum, in addition, the Calsski Harabasz indicator so that the clustering result is more easily determined.

Farahnakian and Heikkonen *et al* [18] designed efficient Intrusion Detection Systems (IDSs) and have gained a lot of attractions due to huge increase different kinds of attacks and network traffic. In this a deep auto-encoder approaches for improving the intrusion detection system. The auto- encoder is one of the most interesting models to extract features from the high-dimensional data in the context of deep learning. Proposed Deep Auto-Encoder based Intrusion Detection System (DAE-IDS) consists of four auto-encoders which the output of the auto-encoder at the current layer is used as the input of the auto-encoder in the next layer. In addition, an auto-encoder at the current layer is trained before the auto-encoder at the next layer. To train DAE-IDS, we utilized a greedy unsupervised layer-wise training mechanism that helps to improve the deep model performance. After training four auto-encoders, here used a softmax layer to classify the inputs into the normal and attack. Here the KDD-CUP'99 dataset to evaluate the performance of DAE-IDS as this dataset has been utilized extensively for evaluating IDSs. The proposed approach achieved detection accuracy 94.71% on the total 10% KDDCUP99 test dataset.

3. Working Methodology

Intrusion Detection system aims to identify malicious or abnormal behaviour on the network and prevent security

breaches. Support Vector Machine (SVM) by building non-linear decision bounds, this process executes regression and classification tasks, in case SVM to perform large classification and regression task due to the nature of the function in various data phase as shown in the fig1.

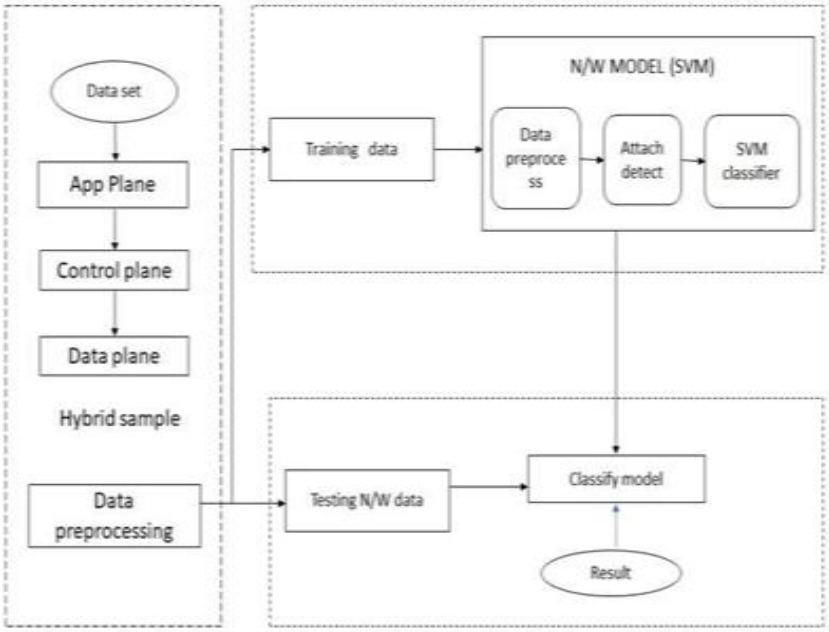

Fig 1: Outline of Proposed System

Machine learning algorithms can be used to analyze large amounts of network data and identify complex relationships between network features that may not be noticeable to a human observer. The algorithms can then be used to detect intrusions in real-time by analyzing new incoming network data. Some commonly used machine learning algorithms for intrusion detection include Random Forest, Support Vector Machine and Gradient Boosting. These algorithms can be trained on labeled data, where the ground truth of normal and anomalous behavior is known, or an unlabeled data, where the algorithm must learn to identify anomalies on its own.

4. Implementation

A. Support Vector Machine (SVM)

Support Vector Machine analyzes the data, defines the decision boundaries and uses the kernels for computation which are performed in input space as shown in fig 2. The input data are two sets of vectors of size m each. Then every data represented as a vector is classified in a particular class. Experimental results show that IDS with an optimized NSL-KDD dataset using the best feature set algorithm based on Information Gain Ratio increases the accuracy of 96.24% and minimizes the false alarm rate. Support Vector Machine has become one of the popular techniques for anomaly intrusion detection due to their good generalization nature and the ability to overcome the curse of dimensionality. Another positive aspect of SVM is that it is useful for finding a global minimum of the actual risk using structural risk minimization, since it can generalize well with kernel tricks even in high-dimensional spaces under little training sample conditions. The SVM can select appropriate setup parameters because it does not depend on traditional empirical risk such as neural networks [19]. One of the main advantages of using SVM for IDS is its speed, as the capability of detecting intrusions in real-time is very important. SVMs can learn a larger set of patterns and be able to scale better, because the classification complexity does not depend on the dimensionality of the feature space. SVMs also have the ability to update the training patterns dynamically whenever there is a new pattern during classification [20].

Input: Training data Xi Labels Yi
Output: Sum of weight vector, α array, b and SV Initialize
$\alpha i = 0$, $fi = -Yi$
Compute *bhigh, lhigh, blow, llow* Update *αihigh* and *αilow* Repeat
Update *fi*
Compute: *bhigh, lhigh, blow, llow*
Update *αihigh* and *αilow* Until *blow* < = *bup*+2 *r* Update the threshold b
Store the new *α1* and *α2* values
Update weight vector w if SVM is linear

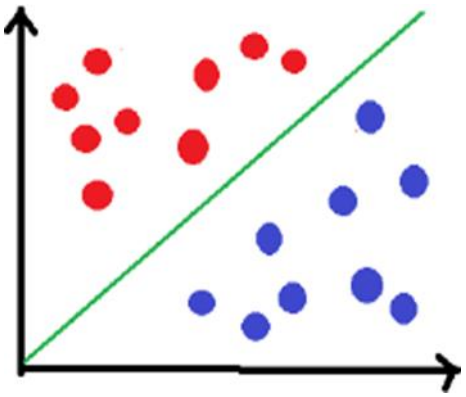

Fig 2: Basic of SVM

B. Random Forest Algorithm (RF)

Random Forest is an ensemble learning algorithm and does not have a single formula. However, its underlying principles can be described as follows:

1. Decision Trees: Random Forest is based on decision trees, which are tree-like structures used to make predictions by

considering a series of decisions based on the values of the input features. Each internal node of the decision tree represents a decision based on a specific feature, while each leaf node represents a prediction as shown in fig 3.

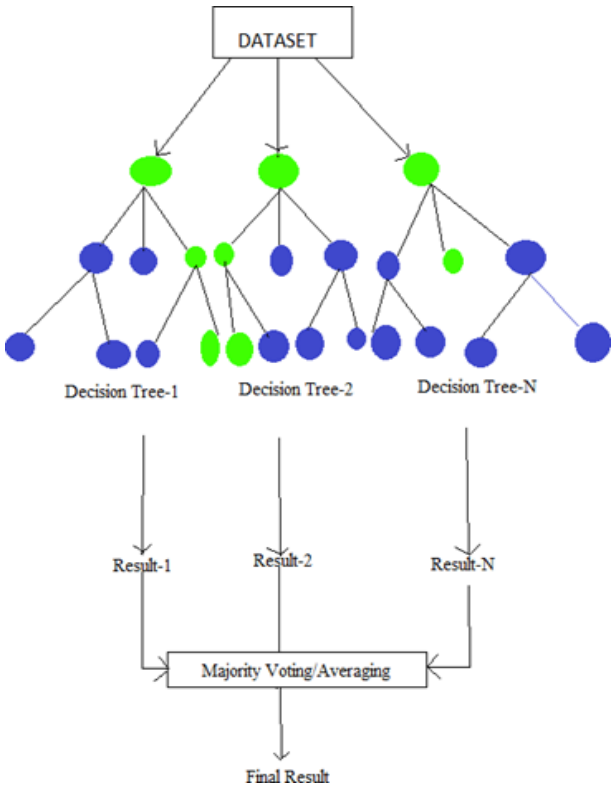

Fig 3

2. Bagging: Random Forest uses an ensemble approach known as bagging (bootstrapped aggregating) to create multiple decision trees and average their predictions. Bagging involves creating multiple bootstrapped samples of the training data and training a decision tree on each sample as shown in fig 4.

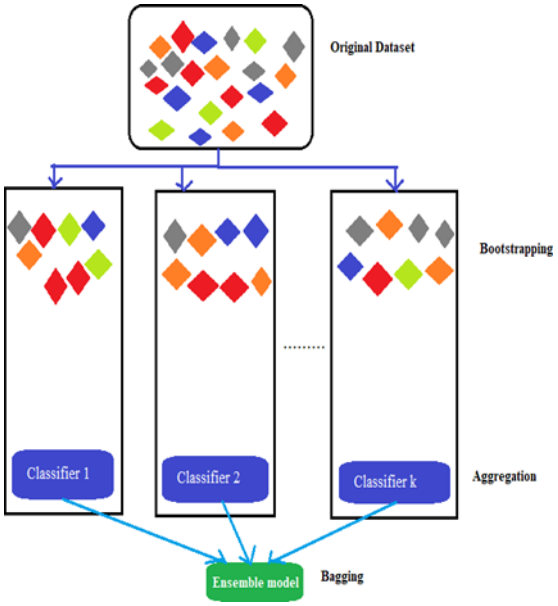

Fig 4: Bagging

3. Random Subspace Method: Random Forest also employs a feature selection method known as the random subspace method. This involves selecting a random subset of the features for each split in each decision tree, rather than using all the features as shown in fig 5.

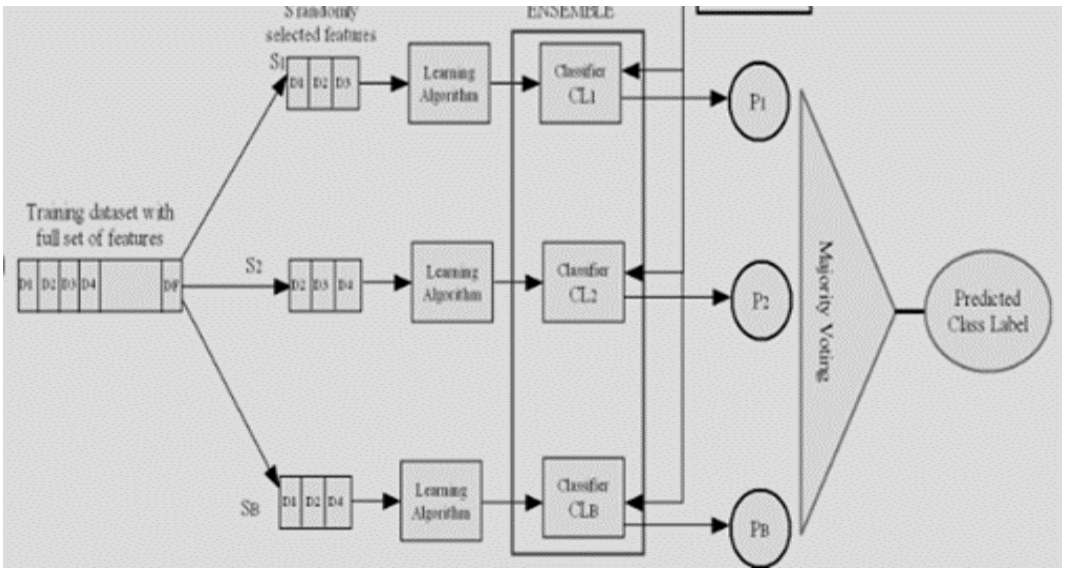

Fig 5: Random Subspace Method

4. Aggregation: The predictions of the individual decision trees are combined to make a final prediction using a majority vote or averaging method as shown in fig 6.

Fig 6: Aggregation

C. Gradient Boosting Algorithm

Gradient Boosting is a machine learning algorithm that does not have a single formula, but instead is based on a series of iterative steps to create an ensemble model is shown in the fig 7. However, some of the key mathematical concepts involved in the algorithm include:

1. Loss Function: The gradient boosting algorithm is based on the idea of minimizing a loss function that measures the difference between the model's predictions and the actual values. Common loss functions used in gradient boosting include mean squared error, logarithmic loss, and exponential loss.

2. Residuals: The residuals are the difference between the model's predictions and the actual values. These residuals are used to train subsequent models in the gradient boosting algorithm.

3. Boosting: Boosting is a method of combining weak models to create a strong model. In gradient boosting, this is achieved by iteratively training new models to fit the residuals of the previous model.

4. Gradient Descent: Gradient descent is a optimization method that is used to minimize the loss function by updating the model parameters in the direction of the negative gradient.

5. Decision Trees: Gradient boosting uses decision trees as the underlying weak model. Decision trees are tree-like structures used to make predictions based on the values of the input features.

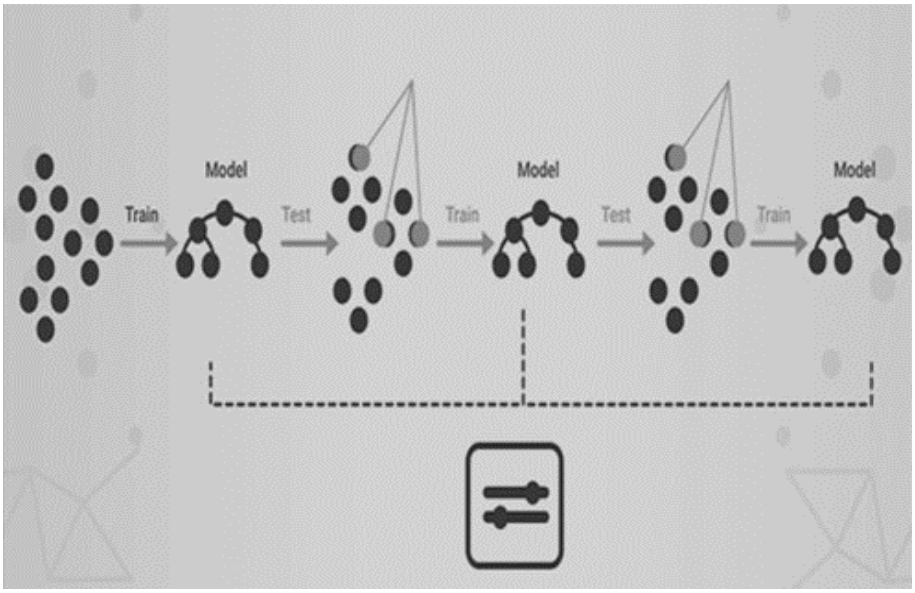

Fig 7: Concept of Gradient boosting algorithm

5. Result Analysis

The below table represents the total detection performance of the SVM model. It is satisfactory and stable also achieves the highest level compared to other methods as represented in table 1.

Table 1

	Normal	DOS	Probe	R2L	U2R
Normal	**9325**	123	217	40	6
DOS	14	**5667**	60	0	0
Probe	131	79	**896**	0	0
R2L	1664	4	9	**522**	0
U2R	29	4	0	2	**2**

The values of the leading diagonal denote the number of correctly classified records of the testing dataset. "R2L" and "U2R" have more samples that are identified incorrectly.

6. Conclusion

The Proposed solution demonstrates how to build an efficient Intrusion Detection System (IDS) by using Support Vector Machine (SVM), Clustering. These Algorithm was found to be the promising in the KDD bench Mark IDS data set. Machine learning Algorithms can be used to analyze large amounts of network data and identify complex relationship between network features that may not be noticeable to a human observer. Also these Algorithms can be trained on labeled data, where the ground truth of normal and anomalous behaviour is known, or on unlabeled data, where the algorithm must learn to identify anomalies on its own.

7. References

1. Dr. S. Vijayarani and Ms. Maria Sylviaa S, Intrusion Detection System-A Study", International Journal of Security, Privacy and Trust Management (IJSPTM) Vol 4, No 1, February 2015 [Online]. Available:https://airccse.org/journal/ijsptm/papers/4115ijsptm04.pdf.
2. S Shin, G Gu. Cloud Watcher: Network security monitoring using OpenFlow in dynamic cloud networks (or: How to provide security monitoring as a service in clouds?), in Proc. IEEE Int. Conf. Netw. Protoc, 2012, 1-6.
3. CJ Chung, P Khatkar, T Xing, J Lee, D Huang. NICE: Network intrusion detection and countermeasure selection in virtual network systems, IEEE Trans. Depend. Secure Comput. 2013; 10(4):198-211.
4. T Xing, Z Xiong, D Huang, D Medhi. SDNIPS: Enabling software-defined networking-based intrusion prevention system in clouds, in Proc. Int. Conf. Netw. Serv. Manage. Workshop, 2014, 308-311.
5. JS Cui, C Guo, L Chen, YN Zhang, D Huang. "Establishing process-level defense-in-depth framework for software defined networks, J Softw. 2014; 25(10):2251-2265. doi: 10.13328/j.cnki.jos.004682.
6. Jawad, Mohammad, Fadhil Tamara. A Signature Best feature selection matching using FAST and genetic algorithm". IOP Conference Series: Materials Science and Engineering. 2020; 870:012132. 10.1088/1757-899X/870/1/012132.
7. GE Hinton, R Salakhutdinov. Reducing the dimensionality of data with neural networks, Science. 2006; 313(5786):504-507. doi: 10.1126/science.1127647.
8. ZM Hira, DF Gillies. A review of feature selection and feature extraction methods applied on microarray data, Adv. Bioinf., 2015, 1-13, 2015, doi: 10.1155/2015/198363.
9. A Kannan, GQ Maguire Jr, A Sharma, P Schoo. Genetic algorithm based feature selection algorithm for effective intrusion detection in cloud networks," in Proc. IEEE 12th Int. Conf. Data Mining Workshops, 2012, 416-423. [Online]. Available: http://dx.doi.org/10.1109/icdmw.2012.56.
10. A Kannan, *et al*. A novel cloud intrusion detection system using feature selection and classification," Int. J. Int. Inf. Technol. 2015; 11(4):1-15. doi: 10.4018/ijiit.2015100101.
11. O Osanaiye, *et al*. Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing, EURASIP J. Wireless Commun. Netw, 2016, 1, 2016, Art. no. 130, doi: 10.1186/s13638-016-0623-3.
12. A Javadpour, S Kazemi Abharian, G Wang. Feature selection and intrusion detection in cloud environment based on machine learning algorithms, in Proc. IEEE Int. Symp. Parallel Distrib. Process. Appl. IEEE Int. Conf. Ubiquitous Comput. Commun, 2017, 4171421. [Online].Available:http://dx.doi.org/10.1109/ispa/iucc.2017.00215.
13. NM Ibrahim, A Zainal. A feature selection technique for cloud IDS using ant colony optimization and decision tree," Adv. Sci. Letts. 2017; 23(9):9163–9169. doi: 10.1166/asl.2017.10045.
14. RC Aygun, AG Yavuz. Network anomaly detection with stochastically improved autoencoder based models," in Proc. IEEE 4th Int. Conf. Cyber Secur. Cloud Comput., 2017, 193-198.
15. Shadi Aljawarneh, Monther Aldwairi, Muneer Bani Yassein, Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model, Journal of Computational Science, Volume 25, 2018, Pages 152-160, ISSN 1877-7503, https://doi.org/10.1016/j.jocs.2017.03.006.
16. Sultana, N., Chilamkurti, N., Peng, W. et al. Survey on SDN based network intrusion detection system using machine learning approaches. Peer-to-Peer Netw. Appl. 2019; 12:493501. https://doi.org/10.1007/s12083-017-0630-0.
17. K Peng, VCM Leung, Q Huang. Clustering Approach Based on Mini Batch Kmeans for Intrusion Detection System Over Big Data in IEEE Access. 2018; 6:11897-11906. doi:10.1109/ACCESS.2018.2810267.Available:https://ieeexplor e.ieee.org/document/8304564
18. F Farahnakian, J Heikkonen. A deep auto-encoder based approach for intrusion detection system, 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Korea (South), 2018, 178-183. doi: 10.23919/ICACT.2018.8323688.
19. Xiaolong Wu, Yirui Wei, Dabao Feng. Jun 2020 Network Attacks Detection Methods Based on Deep Learning Techniques, Available: https://doi.org/10.1155/2020/8872923.
20. Mohsin Abdulazeez, Adnan. Machine Learning Applications based on SVM Classification: A Review, 2021. Available: https://www.researchgate.net/publication/351275238_Machine_Learning_Applications_based_on_SVM_Classificatio n_A_Review.