JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A NOVEL AND EFFICIENT INTRUSION DETECTION STRATEGY USING EFFECTIVE FEATURE SELECTION

S.Vijayalakshmi, Aiswaryaa Velumani, Harini Chokkanathan

Assistant Professor, Student, Student
Department of Information Technology,
Meenakshi Sundararajan Engineering College, Chennai, India

Abstract: Monitoring network traffic for intrusions has always been a difficult undertaking. Intrusion detection systems (IDSs) face significant hurdles because of the variety of network threats. Traditional attack recognition techniques frequently use data mining to discover abnormalities, which has the drawbacks of a high false alarm rate (FAR), a low recognition accuracy (ACC), and a limited capacity for generalization. As Machine learning advancements are opening the door for improving intrusion detection systems, we propose an Intrusion Detection System that uses Feature selection to enhance the comprehensive capabilities of IDS and boost network security. Our system uses a feature selection approach that integrates Principal Component Analysis, a well-liked unsupervised learning method for reducing the dimensionality of data, and Recursive Feature Elimination, which fits a model and eliminates the weakest feature. These two methods make learning more convenient and effective. The KDD and UNSW datasets are input into the system to train and evaluate the Deep Neural Network's capabilities and requirements, respectively. This improves IDS' formalism and clarity for the Internet of Things and Cloud Computing, which are booming in the 5G-ready technological industry right now.

I. INTRODUCTION

Machine learning-based techniques for detecting cyber intrusions have grown in popularity recently. Effective and clever solutions are required since new attacks are becoming more numerous and complicated. Traditional attack recognition techniques typically use data mining to discover anomalies, which has the drawbacks of a high false alarm rate, poor recognition accuracy, and limited generalizability. In addition, detecting unknown attacks in network traffic is the focus of anomaly intrusion detection, making it challenging to do so without assistance from a person. IT managers frequently manually search through system logs to find potential attacks because they find it difficult to keep up with Intrusion Detection System (IDS) notifications. So, our goal was to create an intrusion detection system with improved accuracy and usability.
This documentation's main goal is to offer a thorough introduction to machine learning-based intrusion detection systems. This IDS is built to increase the effectiveness and precision of detecting intrusions across networks. For the selection and omission of features, the method employs RFE and PCA. To train and test the system KDD and UNSW datasets are used respectively with Deep Neural Network.

Recursive Feature Elimination (RFE):

RFE is a feature selection method in the wrapper style that internally also uses filter-based feature selection. RFE attempts to locate a subset of features by successfully eliminating features one at a time until the desired number of features is left, starting with all the features in the training dataset. This is accomplished by first fitting the core model's machine learning algorithm, ranking the features according to relevance, eliminating the least important features, and then re-fitting the model. Up until a certain number of traits are still present, this process is repeated.

Principal Component Analysis (PCA):

Principal component analysis, an unsupervised learning technique, is used in machine learning dimensionality reduction. It is a statistical procedure that converts the observations of correlated characteristics into a set of linearly uncorrelated data using orthogonal transformation. When projecting the high-dimensional data, PCA often seeks out the surface with the lowest dimensionality. The variance of each characteristic is considered by PCA since a high attribute indicates a strong divide between groups, which reduces the dimensionality. Due to the feature extraction technique, it employs only the most important variables that are retained.

Deep Neural Network (DNN):

A deep neural network (DNN) has several hidden layers in between the input and output layers. Complex non-linear relationships can be modelled using DNNs. The primary function of a neural network is to take in a range of inputs, process them through more challenging calculations, and output the outcomes to address real-world issues like categorization. Here, it contains an input layer, where the input features as X_train and Y_train are given. It has hidden layers and an output layer.

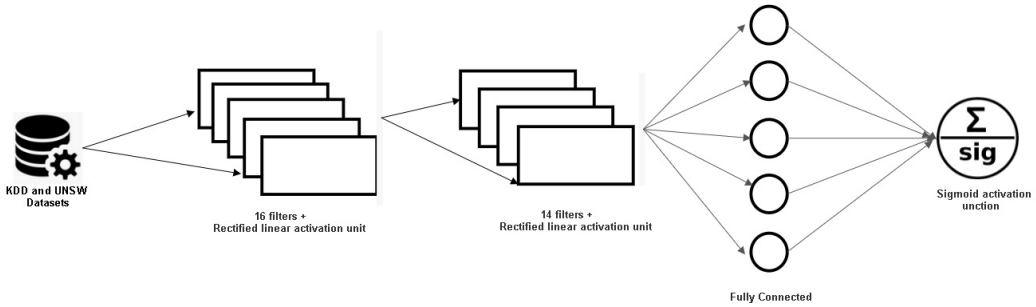

Figure 1 - Working of Deep Neural Network

II. LITERATURE SURVEY

M. A. Siddiqi and W. Pak, "An Agile Approach to Identify Single and Hybrid Normalization for Enhancing Machine Learning-Based Network Intrusion Detection," in IEEE Access, vol. 9, pp. 137494-137513, 2021, doi: 10.1109/ACCESS.2021.3118361. [1]
In this paper, a statistical method is proposed that can identify the most suitable normalization method for the dataset. The normalization method identified by the proposed approach gives the highest accuracy for an intrusion detection system.

G. Pu, L. Wang, J. Shen and F. Dong, "A hybrid unsupervised clustering-based anomaly detection method," in Tsinghua Science and Technology, vol. 26, no. 2, pp. 146-153, April 2021, doi: 10.26599/TST.2019.9010051. [2]
Here, anomaly detection approaches use the normal system activity to build normal-operation profiles, identifying anomalies as behaviors that deviate from the normal ones. Such methods are especially appealing because they are able to potentially detect all known and unknown types of attacks as well as zero-day attacks.

Z. Hu, L. Wang, L. Qi, Y. Li and W. Yang, "A Novel Wireless Network Intrusion Detection Method Based on Adaptive Synthetic Sampling and an Improved Convolutional Neural Network," in IEEE Access, vol. 8, pp. 195741-195751, 2020, doi: 10.1109/ACCESS.2020.3034015. [3]
To ameliorate the comprehensive capabilities of IDS and strengthen network security, they have proposed a novel intrusion detection method based on the adaptive synthetic sampling (ADASYN) algorithm and an improved convolutional neural network (CNN). The standard NSL-KDD dataset is selected to test AS-CNN. The simulation illustrates that the accuracy is 4.60% and 2.79% higher than that of the traditional CNN and RNN models, and the detection rate (DR) increased by 11.34% and 10.27%, respectively.

W. Zhong, N. Yu and C. Ai, "Applying big data based deep learning system to intrusion detection," in Big Data Mining and Analytics, vol. 3, no. 3, pp. 181-195, Sept. 2020, doi: 10.26599/BDMA.2020.9020003. [4]
In order to further enhance the performance of machine learning based IDS, they have proposed the Big Data based Hierarchical Deep Learning System (BDHDLS). BDHDLS utilizes behavioral features and content features to understand both network traffic characteristics and information stored in the payload.

A. Halimaa A. and K. Sundarakantham, "Machine Learning Based Intrusion Detection System," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 916-920, doi: 10.1109/ICOEI.2019.8862784. [5]
Analyzing huge network traffic data is the main work of intrusion detection system. A well-organized classification methodology is required to overcome this issue. This issue is taken in proposed approach. Machine learning techniques like Support Vector Machine (SVM) and Naïve Bayes are applied. The outcomes show that SVM works better than Naïve Bayes. To perform comparative analysis, effective classification methods like Support Vector Machine and Naive Bayes are taken, their accuracy and misclassification rate get calculated.

III. SYSTEM ARCHITECTURE:

The KDD and UNSW Datasets which has various features like protocol__type, root_shell is fed into our system. These has both train data and test data that helps improves detecting variety of intrusions. The intrusion detections are performed as binary classification. The following figure shows the proposed architecture, the dataset is applied feature selection algorithm Recursive Feature Elimination, the number of selected features is 16 then pre-processed with standard scalar, with standardized data, feature reduction technique Principal Component Analysis is performed. The pre-processed data was finally loaded to the DNN model for training and evaluated with two publicly available dataset. To improve the novelty of our project, the effective data pre-processing methods are handled. This helps to build an efficient deep learning model.

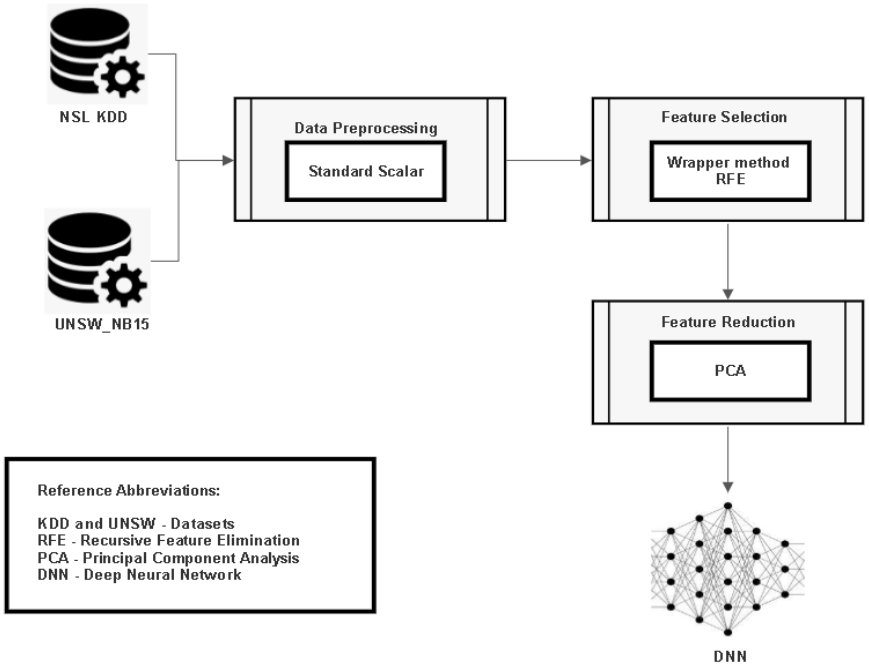

Figure 2 - System Architecture

WORKING:

In this proposed system the intrusion detection technique is implemented with machine learning algorithms. It uses KDD and UNSW datasets which are labelled as normal and abnormal categories; thus, classification algorithm is preferred. The work implements more than two machine learning algorithms to compare and get the best effective model. Recursive Feature Elimination (RFE) is used for selecting the important feature through cross validation. Principal Component Analysis (PCA) is proposed for feature reduction and training. The deep learning algorithm DNN is applied. This novel method is effective in detecting attacks and use feature reduction to perform the most effective learning and avoids over fitting problem.

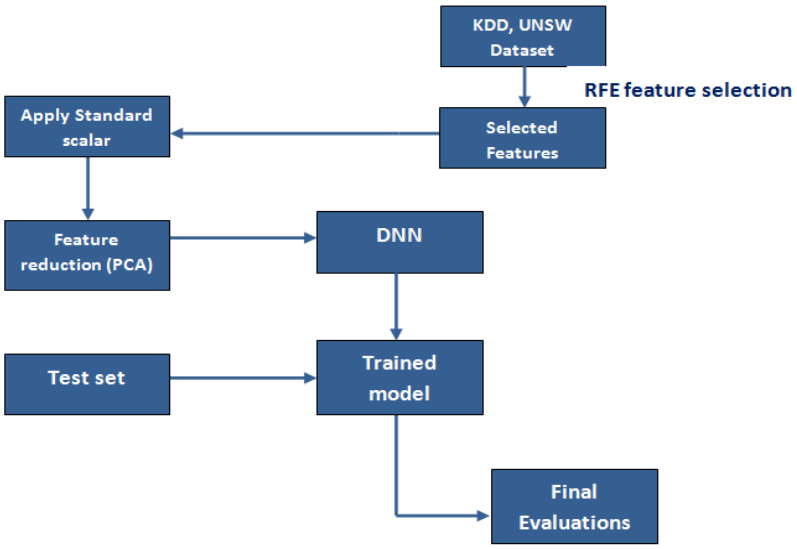

Figure 3 - Overview of the Intrusion detection System

1. Loading the Dataset:

A variety of attacks that can be utilized to gather details about the traits and behaviour of intrusions are present in the KDD-dataset. Data sets frequently include many attributes that might be both damaging to the accuracy of the results and unnecessary. The Internet of Things (IoT) is a network of interconnected devices that generate and consume data by exchanging it back and forth. We eliminate issues such over-fitting, the curse of dimensionality, and imbalance in the dataset using the UNSW-NB15 data set.

2. Feature Selection using RFE:

The scikit-learn, An RFE implementation for machine learning is offered by the Python machine learning library. It is accessible in contemporary libraries. The RFE class in Scikit-Learn provides access to the RFE function. A transform is RFE. The class must first be configured with the chosen algorithm via the "estimator" parameter and the desired number of features via the "n_features_to_select" argument before it can be used. A decision tree or other method of calculating significant scores must be provided by the algorithm.

3. Feature Reduction using PCA:

You must scale the features in your data before performing PCA because scale has an impact on PCA. To assist you standardise the characteristics of the data set onto a unit scale (mean = 0 and variance = 1), use StandardScalar. Conjoining DataFrame along axis results in 1. The last DataFrame before plotting the data is finalDf. You may find out how much information (variance) can be assigned to each of the principal components using the explained variance. This is significant because, although fourdimensional space can be reduced to a two-dimensional space, some of the variance (information) is lost in the process. Using the explained_variance_ratio_ property. PCA is mainly to speed up the fitting of machine learning algorithms.

4. Model Training using DNN:

TensorFlow and PyTorch are the two primary libraries for creating neural networks. TensorFlow was created by Google, and PyTorch was created by Facebook. Both are capable of carrying out the same duties, but the former is better suited for production while the latter is better for creating quick prototypes because it is simpler to master. TensorFlow must first be installed via the terminal. Your machine will translate your Python instructions into CUDA after you set it up, and the GPUs will then process them, making your models run remarkably quickly. We can now import the primary TensorFlow Keras modules into our notebook.

IV. IMPLEMENTATION AND RESULT ANALYSIS

Implementation:

Our work is implemented in Python 3.6.4 with libraries scikit-learn, pandas, matplotlib and other mandatory libraries. We downloaded the dataset from kdd.ics.uci.edu. The data downloaded contains a train set and test set separately with four different classes of intrusions. The train dataset is considered as the train set and the test dataset is considered as the test set. The machine learning algorithm is applied such as decision tree and logistic regression and random forest. We have collected the dataset for the intrusion detection system with the following details from the KDD dataset and we applied machine learning algorithms such as decision tree and regression and random forest.

Result Analysis:

The experiment is conducted on a Deep learning model, Deep neural network (DNN) architecture with two types of datasets KDD and UNSW is trained, evaluated, and tested separately. The experimental result shows that our model achieves around 91.4% of accuracy. We have used 5 epochs, where the accuracy has increased with number of epochs.

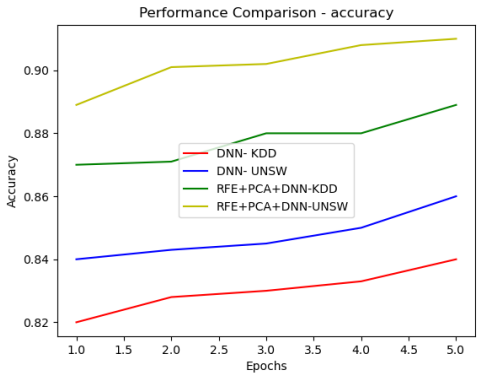

Figure 4

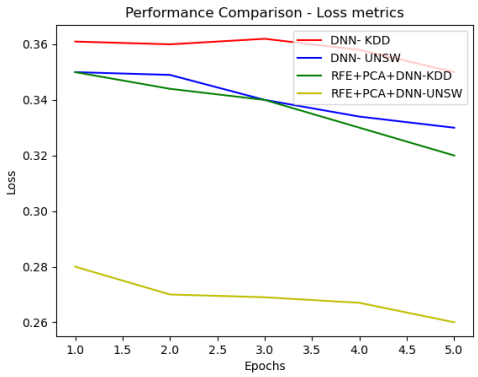

Figure 5

The above figure 4 shows comparison of DNN algorithm accuracy and figure 5 shows comparison of DNN algorithm loss for KDD and UNSW datasets. It is clearly visible that the proposed system outperforms the existing DNN model.

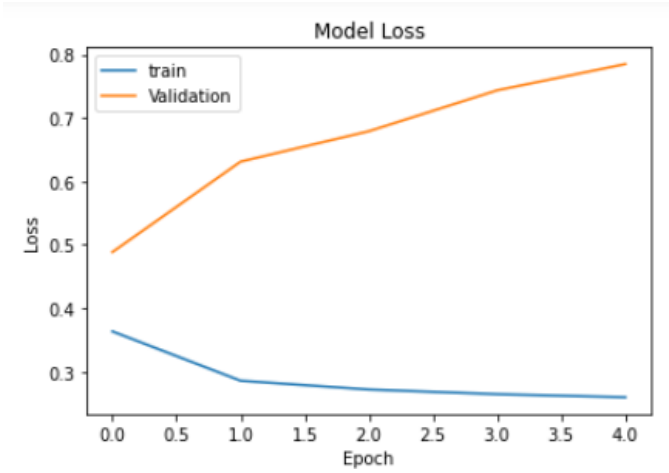

Figure 6 - Model Loss for KDD dataset with DNN algorithm

The above figure 6 shows the model loss registered during the training and validation process of Deep Neural Network algorithm for KDD dataset. The number of epochs is 5 and training loss is reduced to 0.3 at 5 epochs.

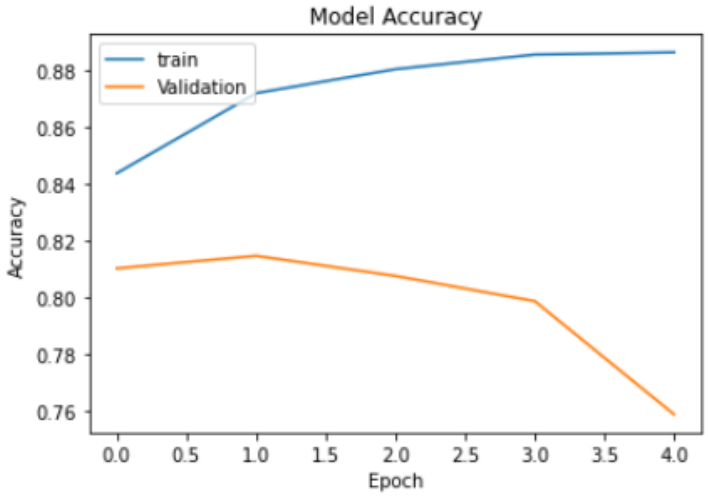

Figure 7 - Model Accuracy for KDD dataset with DNN algorithm

The above figure 7 shows the accuracy of DNN model with KDD dataset. The dataset is trained 5 epochs and the results shows that the accuracy is 88% at epoch 5.

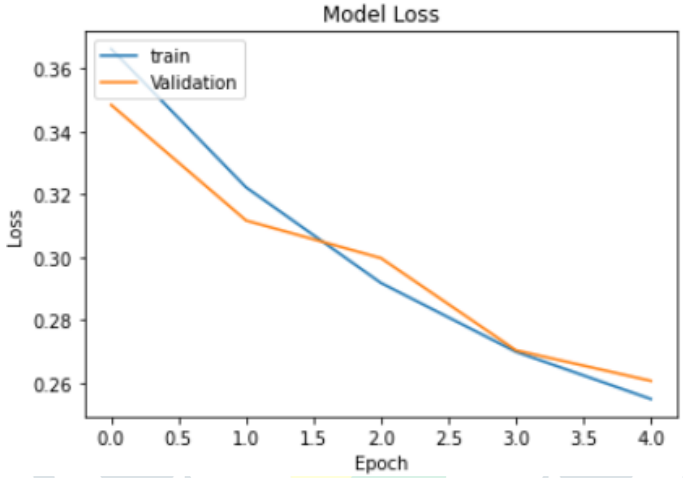

Figure 8 - Model Loss for UNSW dataset with DNN algorithm

The above figure 8 shows the model loss registered during the training and validation process of Deep Neural Network algorithm for UNSW dataset. The number of epochs is 5 and training loss is reduced to 0.26 at 5 epochs.

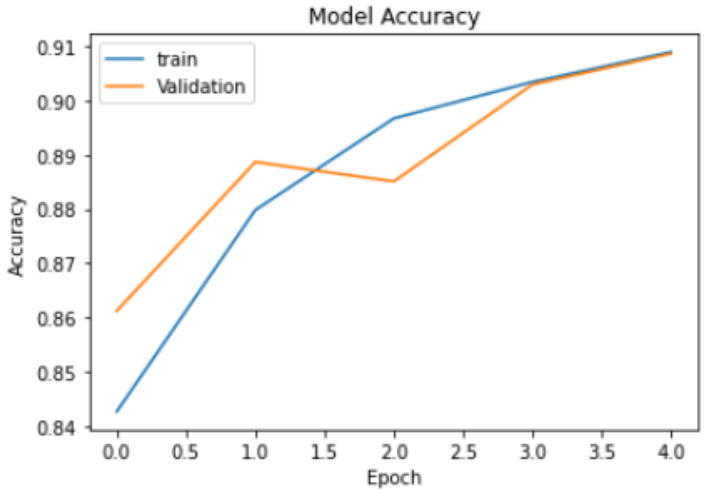

Figure 9 - Model Accuracy for UNSW dataset with DNN algorithm

The above figure 9 shows the accuracy of DNN model with UNSW dataset. The dataset is trained 5 epochs and the results shows that the accuracy is 91% at epoch 5.

Comparison of existing system and proposed work for intrusion detection shown in the below table. The table represents DNN algorithm and proposed feature selected, and feature reduction dataset applied DNN algorithm are compared.

Table 1 - Comparison of DNN and Proposed DNN performance

Algorithm	Dataset	Accuracy	Loss metric
DNN	KDD	84%	0.35
DNN	UNSW	86%	0.33
RFE+PCA+ DNN	KDD	88.9%	0.32
RFE+PCA+ DNN	UNSW	91%	0.26

DNNs can learn complex patterns and relationships in the data by automatically discovering important features during training. They can capture both low-level and high-level representations, making them suitable for a wide range of tasks. But they are computationally intensive and require large amounts of training data. They can be prone to overfitting if the dataset is small or if the model architecture is too complex. RFE is a feature selection technique that iteratively eliminates less relevant features from the input data. When combined with DNNs, RFE can help reduce the dimensionality of the input space, improve model interpretability, and potentially enhance generalization performance by removing noisy or irrelevant features. It contributes to the model's performance when combined with other features such as PCA. PCA is a dimensionality reduction technique that transforms the original features into a new set of uncorrelated variables called principal components. By reducing the dimensionality of the input data, PCA can help improve computational efficiency, mitigate the curse of dimensionality, and potentially enhance generalization performance. Overall, the performance comparison between DNNs alone, DNNs with RFE, and DNNs with PCA is task dependent.

V. CONCLUSION

We aim to address the problems of accuracy and efficiency of detecting intrusions across networks. The network threats include Distributed Denial of Service (DDoS) attacks, user to root (U2R), root to local (R2L) and more. Nowadays the integration of networks for achieving more complex applications is also used like Internet of Things (IoT). These network vulnerabilities create a chance for network attackers to perform illegitimate activities. From a proper analysis of positive points and constraints on the component, it can be concluded that our work has improved the efficiency and accuracy in detecting intrusions. As advances in machine learning pave the path for improving intrusion detection systems, our work has shown to increase network security and broaden the comprehensive capabilities of IDS. Our approach uses a feature selection model that includes principal component analysis, a popular unsupervised learning technique for reducing the dimensionality of data. As a result, feature-based datasets are used to train and evaluate the system. These two algorithms enhance the learning process, making it more practical and efficient. Experimental results showed that accuracy achieved is 91% for UNSW dataset with DNN algorithm.

FUTURE ENHANCEMENTS

- As the further enhancement of intrusion detection system, the CNN based intrusion detection system can be designed with SMOTE-ENN algorithm. It is best suited for intrusion detection on imbalanced network traffic.
- Multiple variants of LSTM like Peephole LSTM, Multiplicative LSTM and Weighted LSTM, in addition to other neural network algorithms and other feature selection algorithms.

REFERENCES

[1] M. A. Siddiqi and W. Pak, "An Agile Approach to Identify Single and Hybrid Normalization for Enhancing Machine Learning-Based Network Intrusion Detection," in IEEE Access, vol. 9, pp. 137494-137513, 2021, doi: 10.1109/ACCESS.2021.3118361.
[2] G. Pu, L. Wang, J. Shen and F. Dong, "A hybrid unsupervised clustering-based anomaly detection method," in Tsinghua Science and Technology, vol. 26, no. 2, pp. 146-153, April 2021, doi: 10.26599/TST.2019.9010051.
[3] Z. Hu, L. Wang, L. Qi, Y. Li and W. Yang, "A Novel Wireless Network Intrusion Detection Method Based on Adaptive Synthetic Sampling and an Improved Convolutional Neural Network," in IEEE Access, vol. 8, pp. 195741-195751, 2020, doi: 10.1109/ACCESS.2020.3034015.
[4] W. Zhong, N. Yu and C. Ai, "Applying big data based deep learning system to intrusion detection," in Big Data Mining and Analytics, vol. 3, no. 3, pp. 181-195, Sept. 2020, doi: 10.26599/BDMA.2020.9020003.
[5] A. Halimaa A. and K. Sundarakantham, "Machine Learning Based Intrusion Detection System," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2019, pp. 916-920, doi: 10.1109/ICOEI.2019.8862784.
[6] I. A. Khan, D. Pi, Z. U. Khan, Y. Hussain and A. Nawaz, "HML-IDS: A Hybrid-Multilevel Anomaly Prediction Approach for Intrusion Detection in SCADA Systems," in IEEE Access, vol. 7, pp. 89507-89521, 2019, doi: 10.1109/ACCESS.2019.2925838.
[7] M. Zeeshan et al., "Protocol-Based Deep Intrusion Detection for DoS and DDoS Attacks Using UNSW-NB15 and Bot-IoT Data-Sets," in IEEE Access, vol. 10, pp. 2269-2283, 2022, doi: 10.1109/ACCESS.2021.3137201.
[8] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," in IEEE Access, vol. 7, pp. 41525-41550, 2019, doi: 10.1109/ACCESS.2019.2895334.
[9] Nguyen Thanh Van, Tran Ngoc Thinh and Le Thanh Sach, "An anomaly-based network intrusion detection system using Deep learning," 2017 International Conference on System Science and Engineering (ICSSE), Ho Chi Minh City, Vietnam, 2017, pp. 210-214, doi: 10.1109/ICSSE.2017.8030867.
[10] A. Alazab, M. Hobbs, J. Abawajy and M. Alazab, "Using feature selection for intrusion detection system," 2012 International Symposium on Communications and Information Technologies (ISCIT), Gold Coast, QLD, Australia, 2012, pp. 296-301, doi: 10.1109/ISCIT.2012.6380910.

[11] K. A. Taher, B. Mohammed Yasin Jisan and M. M. Rahman, "Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection," 2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST), Dhaka, Bangladesh, 2019, pp. 643-646, doi: 10.1109/ICREST.2019.8644161.
[12] M. A. Ambusaidi, X. He, P. Nanda and Z. Tan, "Building an Intrusion Detection System Using a Filter-Based Feature Selection Algorithm," in IEEE Transactions on Computers, vol. 65, no. 10, pp. 2986-2998, 1 Oct. 2016, doi: 10.1109/TC.2016.2519914.