International Research Journal of Education and Technology
Peer Reviewed Journal
ISSN 2581-7795

Sampling Integrated Boosting Classifier for Network Intrusion Detection

A.Sagayapriya[1*], S.Britto Ramesh Kumar[2]
[1] Research Scholar, Department of Computer Science, St.Joseph'sCollege (Autonomous),
Affiliated to Bharathidasan University,Trichy, Tamil Nadu, India.
[2] Associate Professor, Department of Computer Science, St.Joseph's College (Autonomous),
Affiliated to Bharathidasan University, Trichy, Tamil Nadu, India

Abstract

Increase in networking and communication technologies has resulted in improved lifestyle of people. However, the large and sensitive information being transmitted online has become a target for fraudsters. This has resulted in the need for an effective intrusion detection system to safeguard the critical information. This work presents a two phased intrusion detection model, SIBC, that aims to automatically handle data imbalance and also provide effective intrusion detection. Experiments were performed with KDD cup data, NSL-KDD data, and UNSW-NB15 data. Comparisons indicate that the SIBC model performs effective detection of intrusions with accuracy levels with accuracy levels greater than 90% indicating highly effective predictions.

Keywords: Network Intrusion Detection; Ensemble Modeling; Boosting; Sampling; Data Imbalance

1. Introduction

Daily lives of people has experienced a large change due to the increase in communication end networking technologies [1]. Networking technologies are currently very widely used in areas like institutes, industries, educational institutions, shopping end banking. Networking techniques provide several benefits by increasing the quality of life for people [2]. However, they also result in transferring critical information over the network making it vulnerable for attacks. The number of devices connected to the Internet are vastly increasing every day. This results in increased vulnerability the high usage levels has resulted in unauthorized users stealing credible information and misusing them [3]. During the initial stages of networking firewall based detection mechanisms were used. However, the increased networking and communication mechanisms has resulted in the requirement of more complicated intrusion detection mechanisms [4, 5]. Currently, intrusion detection is performed using machine learning models and also deep learning techniques.

From classification perspective, intrusion detection is the process of classifying or differentiating normal traffic from anomalous traffic [6]. The machine learning model is trained using label data. Two approaches to intrusion detection currently exist in the classification domain [7]. They are, signature and anomaly based intrusion detection. Signature based intrusion detection models are trained using the attack signatures. They are highly effective in identifying non attacks [8]. However, in case of unknown attack signatures the model tends to fail. Anomaly based intrusion detection models a train using normal traffic. Variation from normal signatures are considered as anomalies [9]. Further, intrusion detection models are also classified based on their deployment scenarios. The two types of deployment scenarios include host based intrusion detection and network based intrusion detection. The host based intrusion detection models analyze log files from the host system to determine intrusions. Network based intrusion detection models analyze the network packets to detect intrusions [10].

Characteristics of data affects the prediction performance of an intrusion detection system to a large extent. Issues like data imbalance and noise effects the classifiers performance. Qualitative data is very important for a classifier to provide qualitative predictions. This work presents a two phased model that performs both imbalance handling and intrusion detection.

2. Related works

Network intrusion detection has been a major concern for the research community in the last few decades. This section discusses some of the existing recent works in the domain of network intrusion detection.

An evolutionary algorithm based model for effective intrusion detection in networks was proposed by Yerriswamy et al. [11]. This work is based on integrating genetic algorithm and the Grey wolf optimization model. The Grey wolf optimization model has been enhanced to provide better predictions. Further, a new feature selection algorithm has been proposed to identify the effective features for the genetic algorithm. A crow search algorithm model for intrusion detection has been proposed by Theja et al. [12]. This technique is based on opposition based learning. An RNN classifier model is used for the final prediction process. A two phase deep learning model for hybrid intrusion detection has been proposed by Rao et al. [13]. This work uses unsupervised sparse autoencoder in the first stage, to identify the features. The deep neural network model is used in the second stage for predicting and classifying the attack signatures. A similar CNN based model for identifying DO S attacks was proposed by Ngoc et al. [14]. A recurrent neural network based model for both binary and multiclass classification has been proposed by Chuanlong et al. [15], and an LSTM- architecture has been proposed by Kim et al. [16].

A big data based network intrusion detection model for handling imbalanced data has been proposed by Al et al. [17]. This work uses an integration of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) for improved detection of intrusions. Further, imbalance handling is performed using the Synthetic Minority Oversampling Technique (SMOTE) and Tomek-Links sampling methods. An RNN based intrusion detection model for high accuracy classification has been proposed by Chandini [18]. Other similar works using deep learning models for intrusion detection includes CNN based model by Vinayakumar et al. [19], hybrid DNN-kNN model by Souza et al. [20], and a deep learning model by Ding et al. [21].

Extreme learning machine based detection model for identifying network intrusions has been proposed by Alsidani et al. [22]. This work is an integration of Firefly algorithm and fast learning network. The extreme learning machine model has been constructed as a hybrid IDS model for effective intrusion detection. An intrusion detection model based on deep learning architecture has been proposed by Jia et al. [23]. This technique creates a deep belief network that is based on feature entropy. Information gain for each feature is identified and is used for determining redundant features and to reduce the data dimensionality. The information entropy identified from the data is used to determine the level of hidden neurons to be used in the network and also the depth of the network. Imbalance is handled using SMOTE. A kangaroo based intrusion detection system has been proposed by Yazdinejadna et al. [24]. This is a software defined networking based model that aims to handle data plane attacks and malicious behaviors in a networked system. An association rule mining model for detecting network intrusions has been proposed by Lou et al. [25]. It is an adaptive model that uses log files of the cloud computing platform to determine intrusions. A deep learning based model for intrusion detection has been proposed by Yang et al. [26]. It uses a combination of adversarial learning and deep learning do effectively detect known and unknown attacks. The model also concentrates on improving the detection rate of low- frequent attacks.

3. Sampling Integrated Boosting Classifier (SIBC)

Intrusion detection system is a complex mechanism that requires effective distinction of intrusion packets from normal packets in a networked system. The major challenge in performing accurate detection of intrusions is the presence of data imbalance. Data imbalance occurs if one existing class exhibits large number of instances, while the other class exhibits low instance levels. Handling balance during the machine learning process done in two ways. The machine learning model itself is designed to handle imbalance, or a separate module for handling imbalance is added in the architecture. This work uses an independent module to handle data imbalance.

3.1. Data Preprocessing

Data obtained from network transmissions is used as the base data for intrusion detection. The network transmission data contains several additional information like IP address and other user related information. Protocol related information are stored as strings. Machine learning models can only operate on numerical data. Hence, the string data is converted to numerical instance by using one hot encoding. One hot encoding technique creates a new attribute for each of the existing distinct values in the given attribute. The class label is multi class in nature. This label depicts the type of attack. This work uses a binary classification model for prediction. Hence, the multiclass labels are binarized into normal and anomalous traffic. The anomalous traffic is given a label 1 and the normal traffic is given label 0.

3.2. Oversampling for Data Balancing

Network data tends to be imbalanced in nature. The data imbalance is due to the fact that network contains several instances of normal traffic, and very low levels of anomalous traffic. Since intrusions are very rare and occur occasionally, data imbalance occurs. Normal traffic forms the majority class, while the intrusion traffic forms the minority class. Imbalance tends to bias the performance of the classifier. Sampling is the process that is used to counter data imbalance. Two types of sampling techniques are available to handle data imbalance; oversampling, and under sampling.

Undersampling reduces the number of majority class instances to balance with the minority class instances. Random elimination of instances is the mostly used under sampling technique. If the available data has very low number of instances, then reducing the number of majority classes leads to loss of data. This in turn leads to insufficient data for the training process. Oversampling technique increases the number of minority instances to balance with the majority instances. New instances are added by generating data using the existing instances. The number of new instances to be generated is identified by determining the oversampling level.

The oversampling level refers to the number of minority instances to be generated for each majority class instance. The new instances are generated by considering the median values for each attribute of the existing minority instances. The newly generated instances are integrated into the data and the final training data is created.

3.3. Intrusion Detection using Boosting Ensemble

Intrusion detection is performed by using ensemble based modeling technique. Boosting ensemble is created to ensure accurate detection of intrusions. Boosting is the process of repeatedly training the machine learning model by integrating and reducing errors in the machine learning process. Decision tree is used as the base learner for the boosting process.

Let the base learning model be*DT(x)*, and *x* be the training data. The predictions obtained from the model,*y'*is given by equation (1)

$$y' = DT(x) \qquad (1)$$

Errors obtained during the training process is given by equation (2),

$$e = y^{'} - y \qquad (2)$$

Where *y* is the actual class and *y'*is the predicted class.

This process is followed by integrating errors into the model to ensure that the next level predictions are void of these misclassification errors. This is given by equation (3)

$$y^{''} = DT(x) + e \tag{3}$$

This modified learning model also exhibits certain misclassification errors. These are termed as second level errors. They are identified by equation (4)

$$e^{'} = y - y'' \tag{4}$$

These errors are again integrated into the model and next level predictions are identified. The process is repeated until error levels reach the minimum acceptable limits. The model obtained at this stage is used for the test data prediction.

Algorithm-1 SIBC

Input: Imbalanced data (KDD CUP , NSL-KDD, UNSW-NB15)

Output: Predictions on imbalanced data

```
1. Input transaction data
2. Data preprocessing to perform encoding and remove inconsistencies
3. Identify oversampling level (n)
4. For each instance pair i, j in minority class
    a. Identify median for each of the instances
    b. Generate new instance
    c. Integrate it to the training data
    d. Perform this process for n times
5. Create a tree based machine learning model
6. Pass the training data and train the model
7. Identify the errors e
8. Increase the weight of each instance that has been identified as erroneous in the training data
9. If error level less than threshold end training else goto step 6
10. Pass the test data to the trained model to obtain the predictions
```

4. Results and discussion

The proposed SIBC model has been analyzed and its performance has been measured using NSL-KDD, UNSW-NB15 and KDD CUP 99 data. Standard performance metrics like TPR, FPR, TNR, FNR, precision, recall, accuracy, F- measure, and AUC are used as performance analyzers. It could be observed that the SABC model exhibits above average performance in all the given metrics. The performance obtained is shown in table 1.

Table 1: Performance on NSL-KDD, UNSW-NB15 and KDD Cup Dataset

Metric	NSL-KDD	UNSW-NB15	KDD
FPR	0.028	0.009	0.017
TPR	0.882	0.976	0.991
Recall	0.882	0.976	0.991
Precision	0.938	0.976	0.985
TNR	0.972	0.991	0.983
FNR	0.118	0.024	0.009
Accuracy	0.943	0.987	0.987
F-Measure	0.909	0.976	0.988
AUC	0.927	0.983	0.987

Performance analysis in terms of precision and recall is shown in figures 1 and 2. The PR plot shows precision in X axis and recall in Y axis. High values for both precision and recall are expected for a high performing model. The plots for NSL-KDD, UNSW-NB15 and KDD shows that the proposed model exhibits high precision and recall values, greater than 0.8. This shows that the efficiency of prediction of the SIBC model is high irrespective of the data set being used. This generic nature of the model indicates that the SIBC model is capable of exhibiting high performance even on varied data indicating that the model is highly efficient.

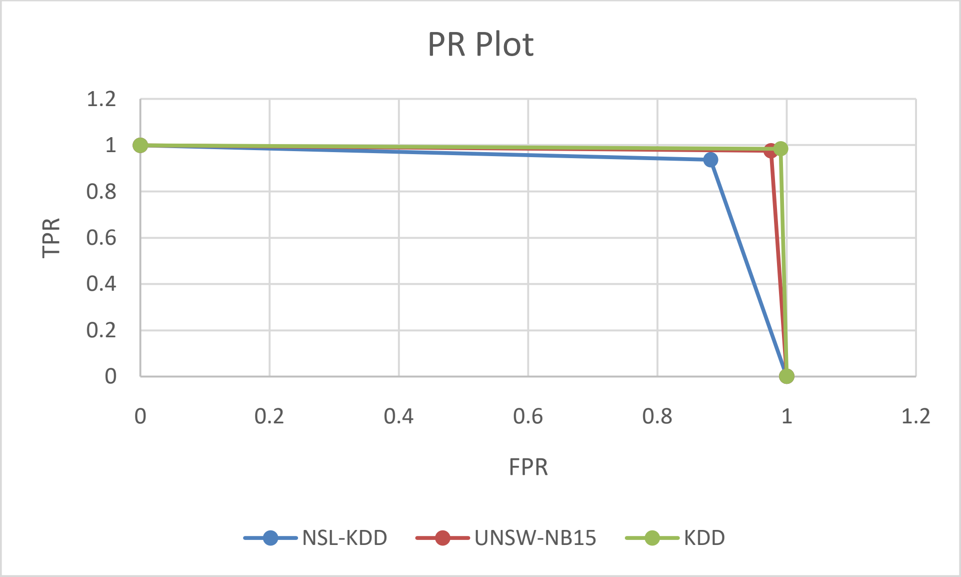

Figure 1: PR Plot

A comparison of the aggregate measures; accuracy, F- measure and AUC has been performed over all the three datasets and the research results are shown in figure 2. The results show that on all the three datasets, the performance of the proposed SIBC model exhibits above average values greater than 90%. This indicates that the SIBC model exhibits very good overall performance in correctly discriminating the intrusion signatures from normal traffic.

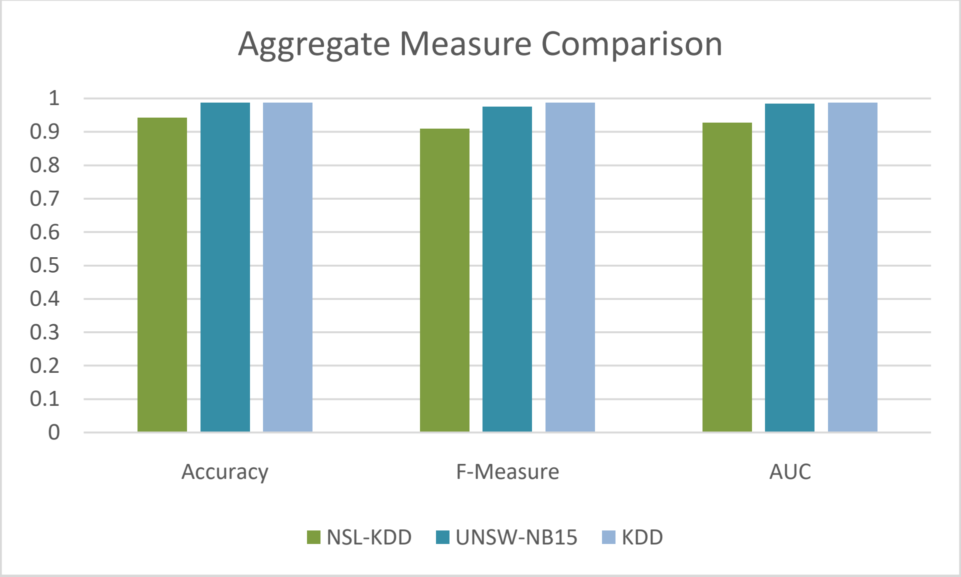

Figure 2: Aggregate Measure Comparison

The proposed SIBC model has been compared with SAVAER-DNN [26] model to comparatively analyze its performance over existing state of the art models. The ROC plot for initial- KDD data set is shown in figure 3. ROC is plotted by considering TPR in Y-axis and FPR in X-axis. Low values of FPR and high values of TPR depict the efficiency of a model. It could be observed from the figure that the proposed SIBC model exhibits low FPR and moderately high TPR. However, the SAVAER- DNN model exhibits higher TPR and also higher FPR. This shows that, although the SAVAER- DNN model exhibits slightly higher TPR, its false alarm levels are much higher compared to the proposed SIBC model.

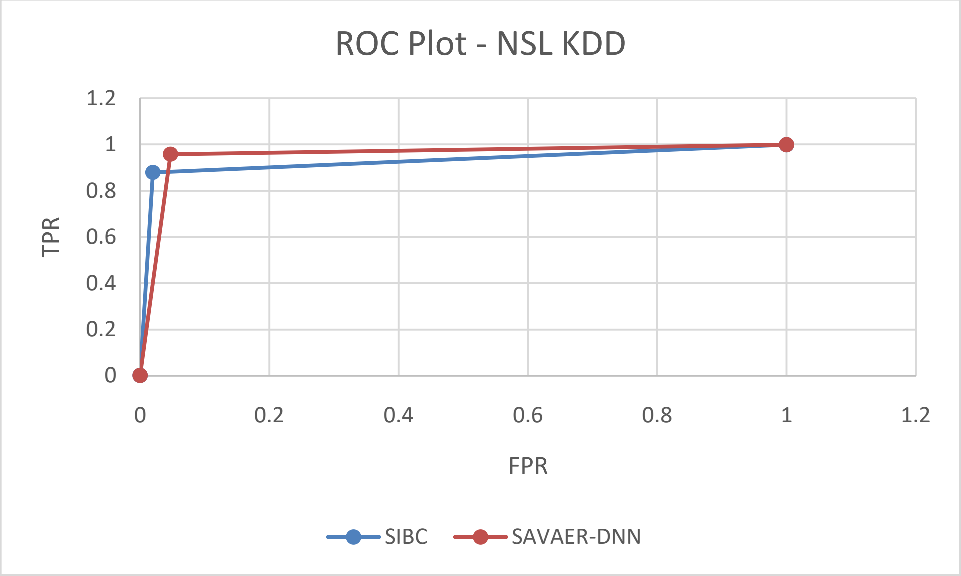

Figure 3: ROC Comparison for NSL-KDD Data

An ROC based comparison over the UNSW-NB15 data set is shown in figure 4. The proposed SIBC model exhibits high TPR and very low FPR levels when compared to the SAVAER-DNN model. This indicates that the SIBC model exhibits improved performance over there UNSW- NB15 dataset.

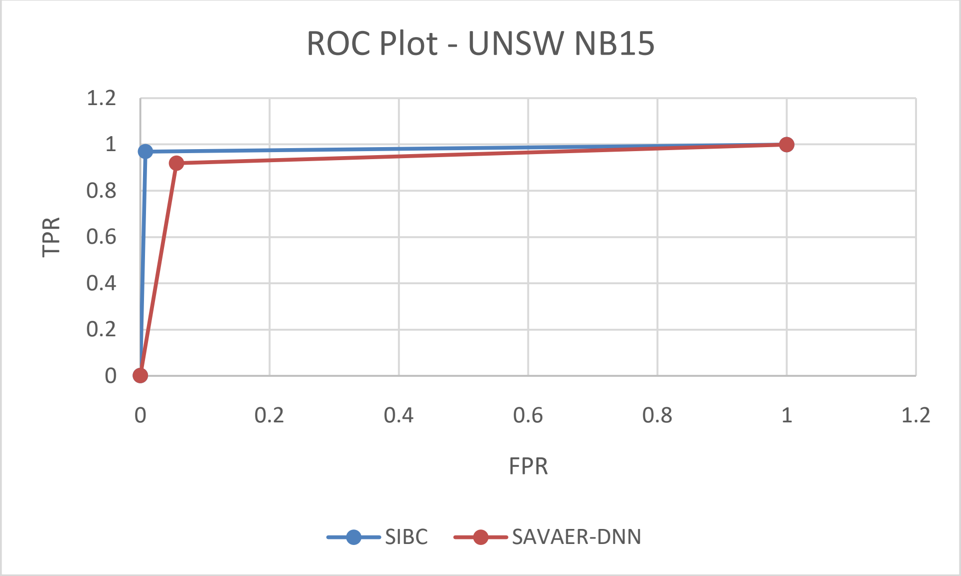

Figure 4: ROC Comparison for UNSW-NB15 Data

A tabulated view of the comparison of FPR, TPR, accuracy and F1 score is shown in tables 2 and 3. The best performances are indicated in bold. Analysis on NSL- KDD data shows that the SIBC model exhibits 2% reduction in false positive levels, and 5% increase in accuracy levels. A 7% reduction in TPR levels can be observed, while both the models exhibit the same F1-Score.

Table 2: Metric Comparison for NSL-KDD Data

Metric	SAVAER-DNN	SIBC
FPR	0.047	**0.02**
TPR	0.959	0.88
Accuracy	0.89	**0.94**
F1-Score	**0.9**	**0.9**

Comparison of metrics on UNSW- NB15 data set is shown in table 3. The comparisonsshow that the SIBC model exhibits high performance on all the metrics. Reduction in false positive levels at 4%, increase in TPR levels at 6%, increase in accuracy levels at 5% and increase in F1- score at 4% were observed from the results.

Table 3: Metric Comparison for UNSW-NB15 Data

Metric	SAVAER-DNN	SIBC
FPR	0.056	**0.01**
TPR	0.919	**0.97**
Accuracy	0.930	**0.98**
F1-Score	0.935	**0.97**

5. Conclusion

Detecting intrusions has become one of the major requirement for the current highly networked scenario. Effectively detecting intrusions aid in better reliability for the users. This work presents an ensemble based model that aids in improving the quality of the data to obtain improved predictions. Performance of the proposed model has been analyzed using three varied datasets. The results show that the proposed SIBC model exhibits highly effective production levels greater than 90%. Comparisons were performed with existing state of the art model is SAVAER-DNN model and the results show that SIBC model exhibits better intrusion detection levels. Limitations of the proposed model is that it exhibits slightly reduced TPR levels in NSL- KDD data set. Future extensions of this work will deal with improving the performance and also to provide multi class classification.

6. References

[1] M. Adnan, M. Rahim, M. Hasanali, K. Al-Jawaheri and K. Neamah, "A Review of Methods for The Image Automatic Annotation", *Journal of Physics: Conference Series*, vol. 1892, no. 1,2021, p012002,doi: 012002.10.1088/1742-6596/1892/1/012002.

[2] H.A.S. Ahmed, M.H. Ali, L.M. Kadhum, M. Fadli, B. Zolkipli, Y.A. Alsariera, “A review of challenges and security risks of cloud computing”, *Journal of Telecommunication, Electronic and Computer Engineering*. vol.9 (1), (2016), pp. 87–91.

[3] KohbalanMoorthy, Mohammed Hasan Ali, MohdArfian Ismail, Chan Weng Howe, MohdSaberiMohamad, SafaaiDeris, “An evaluation of machine learning algorithms for missing values imputation”, International*Journal of Innovative Technology and Exploring Engineering*, vol. 8 (12S2), 2019, pp. 415–420.

[4] M.H. Ali, M.F. Zolkipli, M.M. Jaber, M.A. Mohammed, “Intrusion detection system based on machine learning in cloud computing”, *Journal of Engineering Applied Science*,vol.12 (16), (2017,pp. 4241–4245.

[5] M. H. Ali, M. Fadlizolkipi, A. Firdaus and N. Z. Khidzir, "A hybrid Particle swarm optimization -Extreme Learning Machine approach for Intrusion Detection System," *in IEEE Student Conference on Research and Development (SCOReD)*, 2018, pp. 1-4, doi: 10.1109/SCORED.2018.8711287

[6] K. Scarfone, P. Mell, “*Guide to intrusion detection and prevention systems (IDPS*)”, National Institute of. Standard Technology, Gaithersburg,MD,2007,pp.1-127,doi://dx.doi.org/10.6028/NIST.SP.800-94.

[7] V. Herrera-Semenets, L. Bustio-Martínez, R. Hernández-León and J. van den Berg, "A multi-measure feature selection algorithm for efficacious intrusion detection*", Knowledge-Based Systems*, vol. 227, 2021, p. 107264, 10.1016/j.knosys.2021.107264

[8] Guillermo Francia, LeventErtaul, Luis Hernandez Encinas, Eman El-Sheikh, Kevin Daimi, *Computer and Network Security Essentials*, Springer, 2018, p. 618,doi:doi.org/10.1007/978-3-319-58424-9.

[9] ShadiAljawarneh, MontherAldwairi, MuneerBaniYassein, “Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model”, *Journal of Computer Science*, vol. 25 ,2018,pp.152–160.

[10] C. Song, Y. Sun, G. Han and J. Rodrigues, "Intrusion detection based on hybrid classifiers for smart grid", *Computers & Electrical Engineering*, vol. 93, 2021, p. 107212, doi:10.1016/j.compeleceng.2021.107212

[11] Y. T and G. Murtugudde, "An efficient algorithm for anomaly intrusion detection in a network", *Global Transitions Proceedings*, vol. 2, no. 2,2021, pp. 255-260,doi:10.1016/j.gltp.2021.08.066.

[12] R. SaiSindhuTheja and G. Shyam, "An efficient metaheuristic algorithm based feature selection and recurrent neural network for DoS attack detection in cloud computing environment", *Applied Soft Computing*, vol. 100, 2021, p. 106997,doi: 10.1016/j.asoc.2020.106997.

[13] K. NarayanaRao, K. VenkataRao and P. P.V.G.D., "A hybrid Intrusion Detection System based on Sparse autoencoder and Deep Neural Network", *Computer Communications*, vol. 180, 2021, pp. 77-88,doi: 10.1016/j.comcom.2021.08.026

[14] S. Nguyen, V. Nguyen, J. Choi and K. Kim, "Design and implementation of intrusion detection system using convolutional neural network for DoS detection*", Proceedings of the 2nd International Conference on Machine Learning and Soft Computing - ICMLSC '18*, 2018,pp.34-38,dpi: 10.1145/3184066.3184089.

[15] C. Yin, Y. Zhu, J. Fei and X. He, "A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks," in *IEEE Access*, vol. 5, 2017, pp. 21954-21961, doi: 10.1109/ACCESS.2017.2762418..

[16] J. Kim, J. Kim, H. L. Thi Thu and H. Kim, "Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection," *2016 International Conference on Platform Technology and Service (PlatCon)*, 2016, pp. 1-5, doi: 10.1109/PlatCon.2016.7456805..

[17] S. Al and M. Dener, "STL-HDL: A new hybrid network intrusion detection system for imbalanced dataset on big data environment", *Computers & Security*, vol. 110, 2021, p. 102435, 10.1016/j.cose.2021.102435.

[18] Chandini. S B, "Intrusion Detection using Recurrent Neural Networks", *International Journal for Research in Applied Science and Engineering Technology*, vol. 8, no. 6,2020, pp. 2050-2052, doi: 10.22214/ijraset.2020.6335.

[19] R. Vinayakumar, K. P. Soman and P. Poornachandran, "Applying convolutional neural network for network intrusion detection," *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2017*, pp. 1222-1228, doi: 10.1109/ICACCI.2017.8126009.

[20] C. de Souza, C. Westphall, R. Machado, J. Sobral and G. Vieira, "Hybrid approach to intrusion detection in fog-based IoT environments", *Computer Networks*, vol. 180, 2020, p. 107417, . doi: 10.1016/j.comnet.2020.107417.

[21] Y. Ding and Y. Zhai, "Intrusion Detection System for NSL-KDD Dataset Using Convolutional Neural Networks*", Proceedings of the 2018 2nd International Conference on Computer Science and Artificial Intelligence - CSAI '18*, 2018, pp.11-16, doi: 10.1145/3297156.3297230

[22] M. QahatanAlsudani, S. AbbdalReflish, K. Moorthy and M. Mundher Adnan, "A new hybrid teaching learning based Optimization -Extreme learning Machine model based Intrusion-Detection system", 2021, *Materials Today: Proceedings*, doi: 10.1016/j.matpr.2021.07.015

[23] H. Jia, J. Liu, M. Zhang, X. He and W. Sun, "Network intrusion detection based on IE-DBN model", *Computer Communications*, vol. 178,2021, pp. 131-140, doi: 10.1016/j.comcom.2021.07.016

[24] A. Yazdinejadna, R. Parizi, A. Dehghantanha and M. Khan, "A kangaroo-based intrusion detection system on software-defined networks", *Computer Networks*, vol. 184,2021, p. 107688, doi: 10.1016/j.comnet.2020.107688

[25] P. Lou, G. Lu, X. Jiang, Z. Xiao, J. Hu and J. Yan, "Cyber intrusion detection through association rule mining on multi-source logs", *Applied Intelligence*, vol. 51, no. 6, 2020, pp. 4043-4057, doi: 10.1007/s10489-020-02007-5.

[26] Y. Yang, K. Zheng, B. Wu, Y. Yang and X. Wang, "Network Intrusion Detection Based on Supervised Adversarial Variational Auto-Encoder With Regularization," in IEEE Access, vol. 8, 2020, pp. 42169-42184, doi: 10.1109/ACCESS.2020.2977007

[27] A.SagayaPriya and .S.Britto Ramesh Kumar ,” A STUDY ON VARIOUS TOOLS TO OVERCOME SECURITY ISSUES IN BIG DATA”, International Journal of Scientific Research in Computer Science Applications and Management Studies, Volume 7, Issue 4, July 2018.

[28] A.SagayaPriya, Dr.S.Britto Ramesh Kumar, “Intrusion Detection using Attribute Subset Selector Bagging (ASUB) to Handle Imbalance and Noise”, International Journal of Computer Science and Network Security, Vol.22, No.5, PP.97-102, May.2022

[29] A.SagayaPriya, S.Britto Ramesh Kumar, “Engineering Based On Stacking And Features Handling Data Imbalance With Semi-Supervised Intrusion Detection”, Webology (ISSN: 1735-188X), Volume 18, Number 4, PP. 2456-2467, 2021.

[30] A SagayaPriya, S Britto Ramesh Kumar, “Semi-Supervised Intrusion Detection Based on Stacking and Feature-Engineering to Handle Data Imbalance”, INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY,(ISSN 0974-6846), Volume 15, issue 46, PP. 2548-2554,Dec.2022