A Network Intrusion Detection System based on Extreme Gradient Boosting Technique

Bharat Bhushan Mahor
Dept. of CSE & IT
Ravindranath Tagore University
Bhopal, India

Pratima Gautam
Dept. of CSE & IT
Ravindranath Tagore University
Bhopal, India

Abstract— Popular technologies such as cloud computing the Internet of Things and social networking produce vast volumes of network traffic and data. Intrusion detection systems are therefore essential to track the network and dynamically analyze the incoming traffic. The goal of the intrusion detection system (IDS) is to carry out attack control & to provide desired system security support with intrusion details. The numerous approaches to intrusion detection to predict malicious network traffic have been proposed to date. This paper uses NSLKDD to test intrusion detection machine learning algorithms. Our work aims to examine the theoretical viability of ELM by evaluating the advantages and benefits of ELM. In the last section, we pointed out that the ELM does not degrade the generalization capacity in the expected sense with the proper selection of the activation function. In this study, we begin the analysis in a different direction and demonstrate that random ELM also has some adverse effects. Therefore, by using the Extreme Gradient Boosting Technique we have employed a new technique for machine learning to solve ELM problems.

Keywords— network security, IDS, traffic classification, Intrusion Detection, Malicious traffic, Network

1. INTRODUCTION

The problem of network protection has also been increasingly discussed with the fast growth of the internet. A significant subject in the area of network protection is research on the identification of anomaly within the network. *Intrusion detection systems* are used for network data processing and network anomaly detection. In general, IDS is divided into two categories: signatories and detection systems focused on the anomaly. [1]. Signature-based by IDS [2,3] are designed for the detection of intruders creating an anomaly in character libraries with matching network data like Snort intrusion detection systems[3]. The IDSs are highly detectable, but new attacks can be hard to recognize. – Anomaly-based IDSs build models and perform intrusion detection for ordinary network behavior, depending on whether or not behaviors. Such IDSs are excellent for detecting unknown forms of irregular behavior, but the complete detection rate is poor and the inaccurate alarm rate is high.

Researchers worked hard, trying to use a range of approaches of data extraction & ML on intrusion detection systems to increase the detection rate of IdSs and lower the untrue alarm rate.

However, because of the large number of network data and unequal supply of usual and anomalous behavior, poor detection and high inaccurate alarm rate are problems in the majority of IDSs. The sampling & selection of features with hybrid data optimization data is a strong IDS. The sampling of data shall eliminate outliers in a dataset and decrease the -ve effects on intrusion detection of the unbalanced distribution of data.

IDS is mainly divided into three categories in terms of detection, configuration, and expense. NIDS (Network Intrusion Detection Systems) is installed in many locations within the network to monitor and collect incoming and outgoing network interface traffic. HIDS (Host Intrusion Detection Systems) works on the network's hosts or network computers. The server or computer is incorporated, and the host is often named [4]. The anomaly-based IDS tracks and compares network traffic to an already implemented baseline. [5].

Data mining and the exploration of information have gained tremendous attention in the IT industry. IDS-based data mining can reliably classify user data and even predict results that may be used later DM is used in IDS as a way to extract functions in network data traffic for further research such as forecast analysis. [6]. It is a kind of managed ML algorithm where the classifier is compiled from samples of training data and which is used to forecast unfamiliar classrooms. Here data samples, which are already documented, are used for training. The multi-classification algorithms are constructed by combining two or more.

I. LITERATURE REVIEW

Shen et al. [7] Have two algorithms proposed i.e. Multidimensional assessment (MA) extraction and secondary extraction and sampling of features Secondary feature (SFES). The entire database is separated using the MA algorithm into separate subsets. By evaluating each function individually within different categories, the efficiency of the proposed technique is improved. At the same time, the SFES model minimizes the unbalance and difficulty of the classification algorithm. The classification capacity is usually also promoted here. The main emphasis is on optimizing the classification equilibrium for all groups &also providing better classification efficiency.

Yaseen et al. [8] Presented the multi-level amalgam model by using Support Vector Machine (SVM) (ELM). They also suggested improving SVM and ELM with the K-mean algorithm, so that the existing model of classification is done more effectively, classification education is reduced and IDS performance improved. They used KDDcup99 and achieved the accuracy of 95.75% and 1.87% of the wrong warning score.

Goeschel [9] Suggested a new approach to minimizing false positives and improving IDS performance by putting Support Vector Machines (SVM), Decision Trees and Naïve Bayes composed. At first, SVM will be trained to decide whether the illustration is a regular attack or traffic. The next move is to manage attack-related data and classify the attack using the J48 algorithm. The last stage was the classification and the decision tree of the other unclassified attacks.

Gupta et al.[10] have developed a DM Technique to protect data privacy and integrity. Two methods, i.e. Kmeans clustering and linear regression, have been introduced for data preprocessing using two techniques, namely data transformation & data standardization. Linear regression gives

80 percent precision, while clustering K-means gives 67.5 percent accuracy.

Varma et al. [11] have shown that selected highly important features are needed in the typical IDS Preprocessing phase among the features that are already available. A brief analysis of different selection methods is presented in this paper with a significant emphasis on soft computing methods. The results obtained with the use of soft computing methods such as uneven set philosophy and ACO are much better than the tests and evaluations made with the same IDS function selection algorithms

Li et al. [12] Proposed TCM-KNN, a novel edition raised, i.e. the "KNearest Neighbors Transduced Trust Machines." For anomaly detection, they suggested using the selector of functions KDD cup99 dataset. The Chi-square Function Rating approach was used to assess the most important features. According to the findings, the proposed data set contains all the characteristics (99.48% accuracy and 1,74% false-positive rate) and data set from the selection of the top six main characteristics (99.32 percent accuracy and 2.81 percent inaccurate positive rate).

He et al.[13] have made the issue of learning from imbalanced data apparent. Their success has been tested in an unbalanced learning scenario and they have provided cutting-edge solutions to address the imbalanced learning problem. They explored briefly the potential challenges in this area.

Dhote et al. [14] have provided a summary of three main methods classified among different internet traffic groups. They are classified as port-based approaches; payload-based approaches and statistical methods. In this article, which is narrowly divided into filters, wrappers, and embedded methods, the algorithms for the feature assortment are also described. In conjunction with a brief studying of the feature selection approach used here the advantages and inadequacies of these approaches are also discussed. Applicable to different ML algorithms are functional selection techniques. This leads to the study of current work in this field of science.

Shetty [15] Genetic Algorithm has been established to construct a new KDD Cup99 data set network log header. 21 features of the 41 KDD Cup99 Data Set features were included. New network log headers have been used to produce new data relating to new threats using genetic programming. The findings were compared with two cluster methods, K-mean and Kmedoide, on newly established network data in the sense of precision, detection rates, and faulty positive rates.

Srivastav and Rama Krishna [16] To create a successful interrupt recognition framework, a layered neural system structure is suggested. The structures are linked to current interruption recognition approaches that either uses the neural system or take into account the layered structure. The results show that the device introduced has a better detection rate for intrusions & a lower inaccurate warning rate. They have found KDD cup99.

II. RESEARCH METHODOLOGY

PROBLEM STATEMENT: The findings display that the proposed system has an improved intrusion detection rate and a lower inaccurate alerting rate. They have found KDD cup99In order to overcome this issue, previous IDSs typically use security specialists to examine the dataset & to formulate the DBN algorithm based intrusion detection method. The increase in all information varies quickly, and it has become a lengthy and irritating process for DBN to evaluate and extract attacks or screening rules based on complicated data and vast quantities of networks.

PROPOSED METHODOLOGY: The author proposes an IDS network based on the Extreme Gradient Boost classification in this text (XGBoost classifier). XGBoost's principal benefit is that standardization, as with SVM, etc., is not required. Trees do also well if the knowledge is what I call "lumpy," i.e. not monotonous. Earlier, Extreme Learning Machine was used for the same i. e, but in this technique, there were some drawbacks such as one that the randomness of ELM creates a further ambiguity, both in approximation and learning. The other is that ELM with incorrect activation mechanism also has a generalization degradation phenomenon. Thus, due to the regulatory feature XGBoost was taken into account.

XGBoost (Extreme Gradient Boosting) is a well-known technique for gradient boosting (ensemble) that has improved tree-based efficiency and speed (sequential decision-making trees) algorithm. XGBoost falls into the Ensemble Learning boosting category. The learning ensemble consists of a set of predictors to boost prediction accuracy, which is many models. The technique Boosting helps to correct mistakes created by previous models by applying certain weights to the model. The following models are used.

XGBoost Features

Regularized Learning: The time of controlling help reduce the final weights learned to prevent unnecessary fit. The regularized target appears to choose a model with simple and predictive functions.

Gradient Tree Boosting:

The model of tree ensemble cannot be optimized in Euclidean Space with conventional optimization methods. The model has instead trained additively.

Shrinkage and Column Subsampling:

Two additional technologies are used to avoid overfitting, in addition to the controlled purpose. The first invention is Friedman's shrinkage. Recent weight adjustment scales by a factor η after each tree boost point. Shrinking decreases the effect of each tree and leaves room for future trees to develop the model, similar to the rate of learning for stochastic optimization.

Figure 1 is the schematic illustration of XGB showing the significant steps in the implementation of this research.

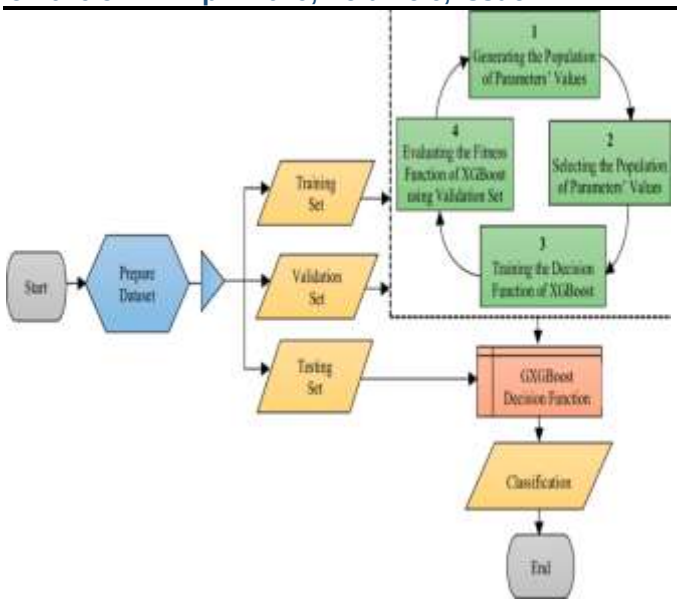

Figure 1: Data Flow diagram of XGBoost

III. SIMULATION RESULTS

The test is carried out with a python-3 program on your desktop. In this experiment, NSL-KDD data was used for training and testing.

```
ELMClassifier(n_hidden=500)

# Compute the error. Results after classification
predictions=elmc.predict(x_test)
accuracy = elmc.score(x_test,y_test)

precision=precision_score(y_test, predictions,average='macro')
recall=recall_score(y_test, predictions,average='macro')

print("Accuracy  : {:.4f}%".format(accuracy*100))
print("Precision : {:.4f}%".format(precision*100))
print("Recall    : {:.4f}%".format(recall*100))
print("--- %s seconds ---" % (time.time() - start_time))

Accuracy  : 94.8603%
Precision : 74.4201%
Recall    : 66.5868%
--- 17.92579221725464 seconds ---
```

Figure 2: Result visualization of SPELM

After loading and assigning a name for each attribute
The dataset is then divided into two sets as one package
Another collection of data as test data. 40% of the KDD data set is used for training data and 60% for research. For training data. After partitioning of the KDD data model, the model is trained with training data for learning purposes.

Figure 2: Result visualization of XGBoost

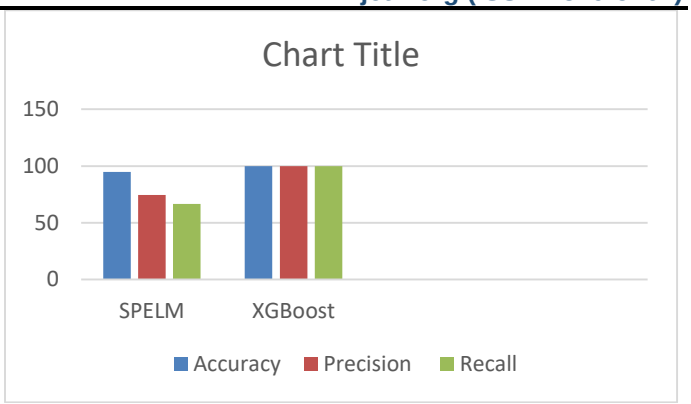

Figure 4: Graphs showing the comparison of SPELM and XGBoost algorithms

.AFTER CALCULATION OF CURRENT ELM OUTPUT INDICATORS MODEL AND PROPOSED MODEL XGB; COMPARISON OF THE TWO MODELS BASED ON THE TOTAL AMOUNT OF TIME AS SHOWN IN FIGURE 5.

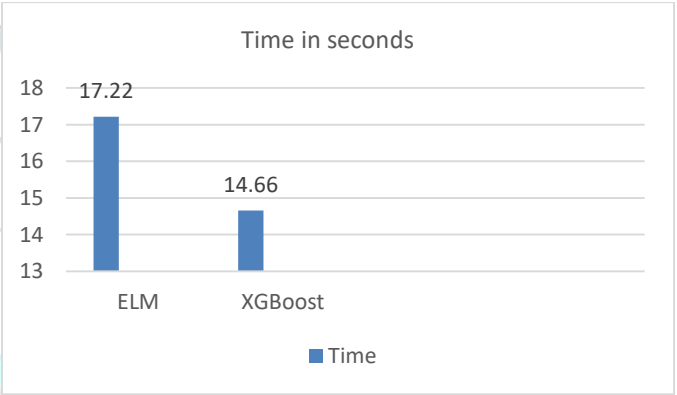

Figure 5 is the total time taken by both the techniques

VI CONCLUSION

Owing to increased numbers of delicate data stored and processed within networking systems, demand for intrusion sensing (IDS) and other safety technologies has increased considerably over the last decade. The test is performed on your desktop using python-3. NSL-KDD data were used for training and testing in this experiment. The researchers have suggested several answers to minimize intrusion effects. By comparing these models with other parameters such as accuracy, accuracy, and reminder, XGBoost Algorithm works better than ELM.

REFERENCES

[1] W. Lee, S. J. Stolfo, and K. W. Mok, "A data mining framework for building intrusion detection models," in Proceedings of the 1999 IEEE Symposium on Security and Privacy, pp. 120–132, USA, May 1999.

[2] H. P. Sasan and M. Sharma, "Intrusion detection using feature selection and machine learning algorithm with misuse detection," International Journal of Computer Science and Information Technologies, vol. 8, no. 1, pp. 17–25, 2016.

[3] J. E. Díaz-Verdejo, P. García-Teodoro, P. Muñoz, G. Maciá-Fernández, and F. De Toro, "A Snort-based approach for the development and deployment of hybrid IDS," IEEE Latin America Transactions, vol. 5, no. 6, pp. 386–392, 2007.

[4] P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli, "A Detailed Investigation and Analysis of using Machine Learning Techniques for Intrusion Detection," IEEE Communications Surveys & Tutorials, pp. 1–1, 2018.

[5] N. V. Patel, N. M. Patel, and C. Kleopa, "OpenAppID - application identification framework next generation of firewalls," International Conference on Green Engineering and Technologies (IC-GET), pp. 1–5, 2016.

[6] V. Bontupalli and T. M. Taha, "Comprehensive survey on intrusion detection on various hardware and software," National Aerospace and Electronics Conference (NAECON), pp. 267–272, 2015.

[7] J. Shen, J. Xia, Y. Shan, and Z. Wei, "Classification model for imbalanced traffic data based on secondary feature extraction," IET Communications: IET Journals, vol. 11, no. 11, pp. 1725–1731, 2017.

[8] W. L. Al-Yaseen, Z. A. Othman, and M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system," Expert Systems with Applications, vol. 67, pp. 296–303, 2017.

[9] K. Goeschel, "Reducing false positives in intrusion detection systems using data-mining techniques utilizing support vector machines, decision trees, and Naive Bayes for off-line analysis," SoutheastCon 2016: IEEE, pp. 1–6, 2016.

[10] D. Gupta, S. Singhal, S. Malik, and A. Singh, "Network intrusion detection system using various data mining techniques," International Conference on Research Advances in Integrated Navigation Systems (RAINS): IEEE, pp. 1–6, 2016.

[11] Manishaben Jaiswal," CLOUD COMPUTING AND INFRASTRUCTURE", IJRAR - International Journal of Research and Analytical Reviews (IJRAR), E-ISSN 2348-1269, P- ISSN 2349-5138, Volume.4, Issue 2, Page No pp.742-746, June 2017,

DOI Member: 10.6084/m9.doi.one.IJRAR19D1251

Available at http://www.ijrar.org/viewfull.php?&p_id=IJRAR19D1251

[12] Manishaben Jaiswal, "COMPUTER VIRUSES: PRINCIPLES OF EXERTION, OCCURRENCE AND AWARENESS ", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.5, Issue 4, pp.648-651, December 2017, http://doi.one/10.1729/Journal.23273 Available at http://www.ijcrt.org/viewfull.php?&p_id=IJCRT1133396

[13] Manishaben Jaiswal "Big Data concept and imposts in business" International Journal of Advanced and Innovative Research (IJAIR) ISSN: 2278-7844, volume-7, Issue- 4, April 2018 available at: http://ijairjournal.in/Ijair_T18.pdf

[14] Manishaben Jaiswal " SOFTWARE QUALITY TESTING " International Journal of Informative & Futuristic Research (IJIFR) , ISSN: 2347-1697 , Volume 6, issue -2 , pp. 114-119 ,October-2018 Available at: http://ijifr.com/pdfsave/23-12-2019214IJIFR-V6-E2-23%20%20OCTOBER%202018%20a2%20files%20mergeda.pdf

[15] P. Ravi KiranVarma, V. ValliKumari, and S. Srinivas Kumar, "A Survey of Feature Selection Techniques in Intrusion Detection System: A Soft Computing Perspective," in Advances in Intelligent Systems and Computing, Singapore: Springer Singapore, vol. 710, pp. 785–793, 2018.

[16] Y. Li, B. Fang, L. Guo, and Y. Chen, "Network anomaly detection based on TCM-KNN algorithm," 2nd ACM symposium on Information, no. 6, pp. 13-19, 2007.

[17] H. He and E. A. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263–1284, 2009.

[18] Y. Dhote, S. Agrawal, and A. J. Deen, "A Survey on Feature Selection Techniques for Internet Traffic Classification," International Conference on Computational Intelligence and Communication Networks (CICN): IEEE, pp. 1375–1380, 2015.

[19] N. P. Shetty, "Using clustering to capture attackers," International Conference on Inventive Computation Technologies (ICICT): IEEE, vol. 3, pp. 1–5, 2016.

[20] N. Srivastav and R. K. Challa, "Novel intrusion detection system integrating layered framework with neural network," in Proceedings of the 2013 3rd IEEE International Advance Computing Conference (IACC), vol. 35, no. 2, pp. 682–689, 2013