A Model for Intrusion Detection in Cybersecurity Using Random Forest Algorithm

Adeniji Oluwashola David[1] and Adeniji Adefolaju O.[2]

[1,2]Computer Science Department,

University of Ibadan,

Ibadan,

Nigeria.

Email: [1]sholaniji@yahoo.com,od.adeniji@ui.edu.ng , [2]adefolajuadeniji@gmail.com

ABSTRACT

Network Intrusion Detection Systems (NIDSs) are essential tools for the network system administrators to detect various security breaches inside an organization's network. A NIDS monitors and analyzes the network traffic to and from the network devices of an organization and raises alarms if an intrusion is observed. The feature and the parameter passed into algorithm may reduce performance of the algorithm. The aim of the research is to develop a model in Cybersecurity using Random forest Algorithm. Machine learning approach was used as a tool for the developed model. The objective of the study is to select the best feature using Principal Component Analysis and tune the parameter using Sequential Mode Based Optimization. There were two data set used for the research. The test+ and the test-21 from NSL-KDD data set. The test+ dataset for the developed model shows 78% Accuracy.79% for Precision and 79% for Recall while for the existing model the accuracy was 88.5%, precision 82.6% and 96.2% Recall. In a similar result for test-21 dataset, the developed model shows 45% Accuracy, 55% Precision and 88.6% for Recall while in existing model 77%, 88.5% and 78.5% for Accuracy, Precision and Recall was gathered respectively. The developed model shows a low classification of 16% for normal packet and 84% classification for anomalies as seen in the confusion matrices. The model was able to identify anomalies with 84% Accuracy which is very good considering the focus on detecting suspicious packet.

.

Keywords: *Intrusion Detection, Random Forest, Hyperparameter Tuning, PCA, Optimization*

African Journal of Computing & ICT Reference Format

Adeniji Oluwashola David and Adeniji Adefolaju O. (2021),
A Model for Intrusion Detection in Cybersecurity using Random Forest Algorithm,
Afr. J. Comp. & ICT, Vol. 14, No. 2, pp. 46 –51

1. INTRODUCTION

Intrusion Detection Systems (IDS) are security tools that, like other measures such as antivirus software, firewalls and access control schemes, are intended to strengthen the security of information and communication systems [1]. The goal of intrusion detection systems (IDS) is to identify attacks from intruders with high accuracy in order to secure internal networks [2]. Nowadays, cyber security is a challenging issue in the cyberspace and it has been increasing dramatically depending on computerization on different application domains including finances, industry, medical, and many other important areas. There is a strong demand for effective Intrusion Detection System (IDS) that is designed to interpret intrusion attempts of incoming network traffic efficiently, intelligently and energy effectively. In order to protect cyber-attack, awareness of an attack is essential to being able to react or defend against attackers. For instant, it is important to know immediately for additional precaution or possibly take law enforcement or legal actions, if information of credit card data has already been stolen. Intrusion detection can also be applied beyond detecting cyber-attack in noticing abnormal system behavior to identify accidents or unexpected conditions. For examples, IDS can be informed anomalies where a human error or malfunction is causing customer credit card numbers to be incorrectly changed several times [3].

Network Intrusion Detection Systems (NIDSs) are essential tools for the network system administrators to detect various security breaches inside an organization's network. An NIDS monitors and analyzes the network traffic entering into or exiting from the network devices of an organization and raises alarms if an intrusion is observed, based on the methods of intrusion detection [4]. The two basic kinds of NIDSs are Signature based Detection (SNIDS) and Anomaly based Detection (ANIDS). In SNIDS, Attack signatures are pre-programmed into the system and pattern matching is then done on incoming traffic, once there is a match the network packet is flagged as an intrusion whereas in ANIDS, packets are flagged as intrusion when they possess attributes that deviate from the normal traffic pattern and this research focuses on the Anomaly Based approach.

2. RELATED WORK

In this section the study describes some of the work that has been done in the area of Network Intrusion Detection. The review in [5] developed a combining classifier model based on tree-based algorithms for network intrusion detection. The NSL-KDD dataset, a much-improved version of the original KDDCUP'99 dataset, was used to evaluate the performance of our detection algorithm. The task of our detection algorithm was to classify whether the incoming network traffics are normal or an attack, based on 41 features describing every pattern of network traffic. The detection accuracy of 89.24 % was achieved using the combination of random tree and NBTree algorithms based on the sum rule scheme, out-performing the individual random tree algorithm. This result represents the highest result achieved so far using the complete NSL-KDD dataset.

Deep belief neural (DBN) networks proved to be the most influential deep neural nets and generative neural networks that stack Restricted Boltzmann Machines. The information in [6] explored the capabilities of DBN's performing intrusion detection through series of experiments after training it with NSL-KDD dataset. The trained DBN network identifies any kind of unknown attack in dataset supplied to it. The model not only detected attacks but also classify them in five groups with the accuracy of identifying and classifying network activity based on limited, incomplete, and nonlinear data sources. The model achieved detection accuracy of about 97.5% for only fifty iterations that is state of art performance. The classification approach in [7] that hybridizes statistical techniques and SOM for network anomaly detection. Thus, while Principal Component Analysis (PCA) and Fisher Discriminant Ratio (FDR) have been considered for feature selection and noise removal, Probabilistic Self-Organizing Maps (PSOM) aim to model the feature space and enable distinguishing between normal and anomalous connections. The detection capabilities of the model can be modified without retraining the map, but only by modifying the unit activation probabilities. This deals with fast implementations of Intrusion Detection Systems (IDS) necessary to cope with current link bandwidths.

Deep learning approach was presented in [8] for flow-based anomaly detection in a Software Defined Networking (SDN) environment. They built a Deep Neural Network (DNN) model for an intrusion detection system and train the model with the NSL- KDD Dataset. In this work, they just use six basic features (that can be easily obtained in an SDN environment) taken from the forty- one features of NSL-KDD Dataset. Through experiments, we confirm that the deep learning approach shows strong potential to be used for flow-based anomaly detection in SDN environments. The review in [9] proposed an anomaly-based network intrusion detection system. The system could analyze huge datasets in a short period of time. They utilized 90.9 GB of a real network packet dataset provided by the Information Security Centre of Excellence at the University of New Brunswick. The system analyzed the packet captured files of this dataset in the environment by using Apache Hadoop and Spark. An approach to implement the system is based on Hive SQL and unsupervised learning algorithms. The accuracy of the detection system is 86.2% with 13% of the false positive rate. These results are promising to detect attacks in real-time. Developed a model called NNET NSA (Neural Network

Negative Selection Algorithm) to evaluate error rate in immune inspired concept using neural network for intrusion detection was presented in [10]. NSLKDDCup1999 dataset was used to test the model.

The results from the developed model show that the model NNET NSA achieved Receiver Operating Characteristics (ROC) showing 90% Area under the Curve (AUC) proportion of accuracy in detection of cyber-crime. The Error rate evaluation of NNET NSA classification of cyber-crime detection was the less by 0.05%, naïve Bayes by 0.16% and SVM by 0.22%, respectively on the R console. Further information on an intelligentsias-scammer filter mechanism using bayesian techniques can be found in [11]. The significant roles of encryption algorithms are numerous and essential in information security. In [12], recurrent neural network algorithm was used for the analysis of the data set at different levels of granularity. And in [13], a Zero Day attack Prediction was carried out. The internet service driven network is a new approach to the provision of network computing that concentrates on the services that are to be provided, as adopted in [14]. The tradeoff between the two protocols can provide a significant impact on the networks.in [15].However there are no adequate provision for quality of service (QoS) in OpenFlow using Flow Label to reduce bits required as a field to match packets in internet protocol six (IPv6)[16].

3. METHODOLOGY

The specific problem associated with intrusion detection of network in cybersecurity is the techniques in identifying anomalies of suspicious packet. There are four stages involved during the experiment which are as follows:

In Stage 1: A NIDS reads all inbound packets and searches for any suspicious patterns. When threats are discovered, based on its severity, the system takes action such as notifying administrators, or barring the source IP address from accessing the network. A part of the model is the module that classifies the inbound packets as an intrusion or not based on the attributes of the packet. The Fig 1 below show an overview of the methodology.

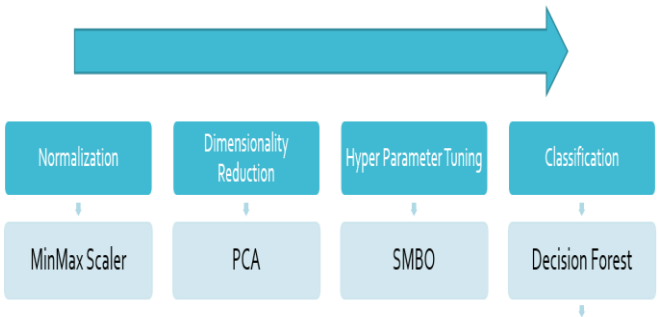


Figure 1: Model Overview.

The Stage 2: make used of dataset in the NSL-KDD dataset described by the 41 features. The Training data contains 125,973 while the Test Data named Test+ contains 22,543 data point. A third Dataset named Test-21 was derived from the Test Data for further testing which contains only 11850 with a classification difficulty level of 21.

Stage 3 involve the Preprocessing which was carried out on the data, after the data has been normalized; the number of features was reduced using Principal Component Analysis (PCA), this gives allowance for much information to be preserved while the number of variables of dataset is reduced.

Stage 4: An effective classifier algorithm; Random forest was implemented for the classification of the packet. Random forest like its name implies, consist of a large number of individual decision trees. Each individual tree in random forest spits out a class prediction and the class with the most votes becomes the mode's prediction. To improve the accuracy and to ensure optimal result of the selected classifier, hyper parameter tuning was employed using Sequential Model Based Optimization (SMBO). The pseudo code used for Random Forest Algorithm is shown below.

1. Randomly select **"k"** features from total **"m"** features.
    1. Where **k << m**
2. Among the **"k"** features, calculate the node **"d"** using the best split point.
3. Split the node into **daughter nodes** using the **best split**.
4. Repeat **1 to 3** steps until "l" number of nodes has been reached.
5. Build forest by repeating steps **1 to 4** for "n" number times to create **"n" number of trees**.

4. RESULT AND DISCUSSION

Before Classification, as described in the methodology, the study implements the Sequential Model Based Optimization on a random forest classifier so as determine the best possible combinations of hyper parameters. The figure 2 below shows the Loss for 30 iterations of the algorithm with the 9th iteration having the least Loss. Hence, the study selects the hyper parameter combination used in the 9th Iterations.

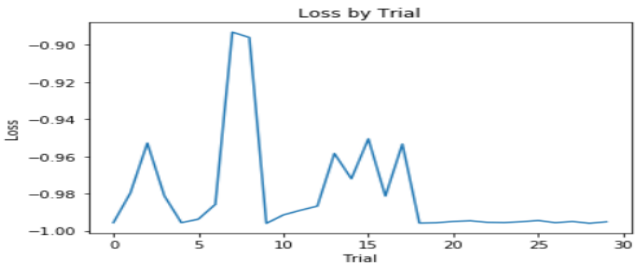


Figure 2. Sequential Mode Based Optimization Trials

The Cross-Validation Accuracy Score for the Train Data Set was a near perfect 99% in the model. The model performed well in classifying normal packets or anomalies however it struggled greatly to classify Normal packets as seen in Test+ and Tes+21 Confusion Matrix in figure 3 and figure 4.

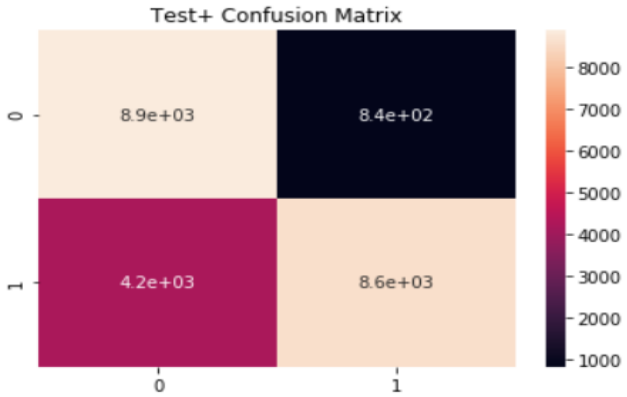


Figure 3: Confusion Matrix for TEST+

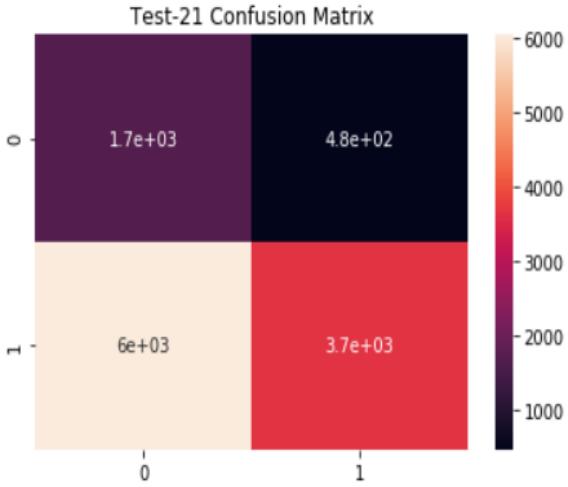


Figure 4: Confusion Matrix for TEST+21

The results of Confusion Matrices in figure 3 shows the binary matrix obtained by the system. The result portrays a instance value of 8.9e+03 normal and 4.2e+03 anomalies packets for the Test+ while in figure 4 the binary instance result for Test21+ shows 1.7e+03 normal and 6e+03 anomalies packets.

The Area under the Curve (AUC) Test+ and Tes+21 AUC Confusion Matrices are shown in figure 5 and figure 6. The Area under the Curve (AUC) provides the proportion of accuracy in detection of cyber-crime.

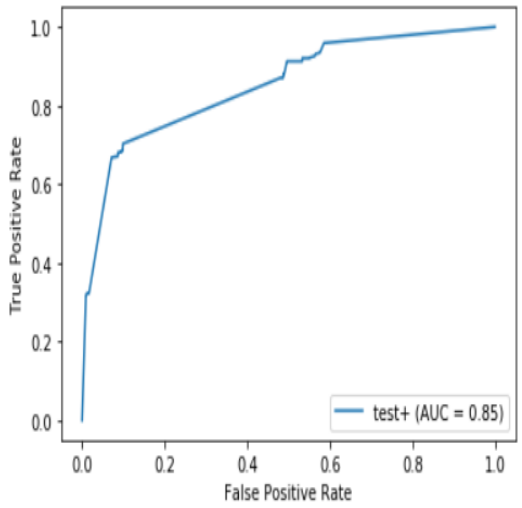
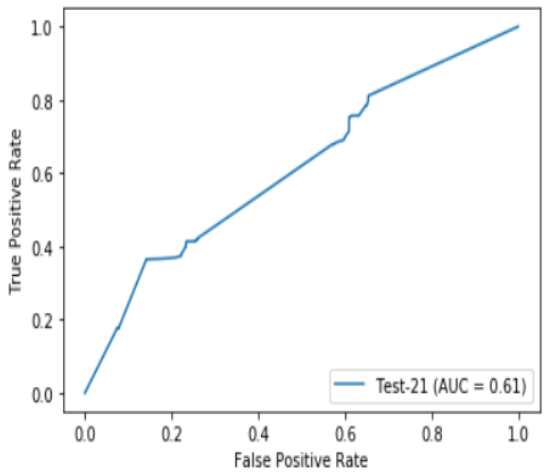


Figure 5: AUC for Confusion Matrix for TEST+

The Area under the Curve (AUC) Test+ provide the result of 85% while Area under the Curve (AUC) Test+21 gives 61% .This shows proportion of accuracy in detection of cyber-crime committed.



Figre 6: AUC for Confusion Matrix for TEST-21

The figures 7 below shows the result of the accuracy of the experiment.

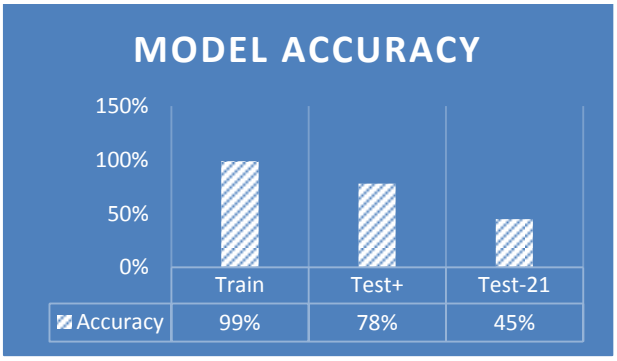


Figure7: Evaluation (Model Accuracy)

In a related simulation for test+ dataset in figure 8 for the developed model shows 78% Accuracy. 79% for Precision and 79% for Recall while for the exiting model the accuracy was 88.5%, precision 82.6% and 96.2% Recall.

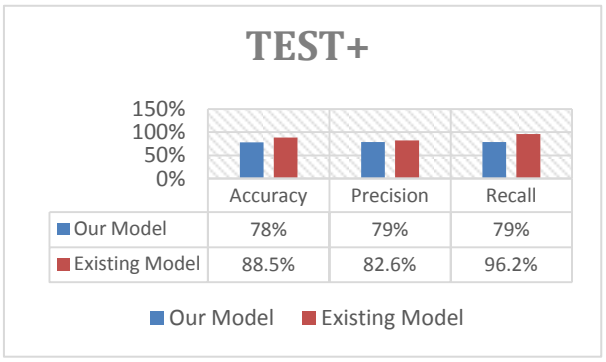


Figure 8: Evaluation (Test+ Dataset)

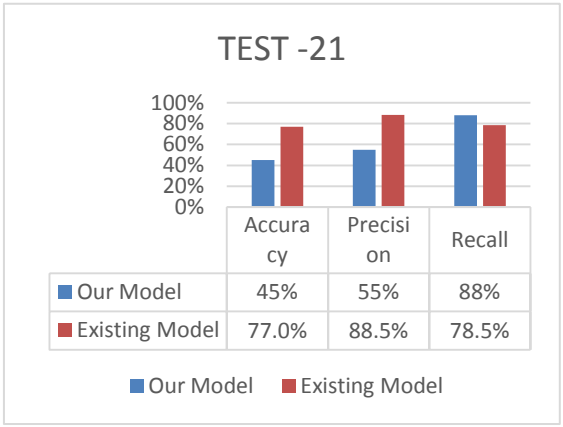


Figure 9: Evaluation (Test-21 Dataset)

In a similar result for test+21 dataset in figure 9 above, the our developed model shows 45% Accuracy, 55% Precision and 88.6% for Recall while in exiting model we have 77%, 88.5% and 78.5% for Accuracy, Precision and Recall was gathered respectively.

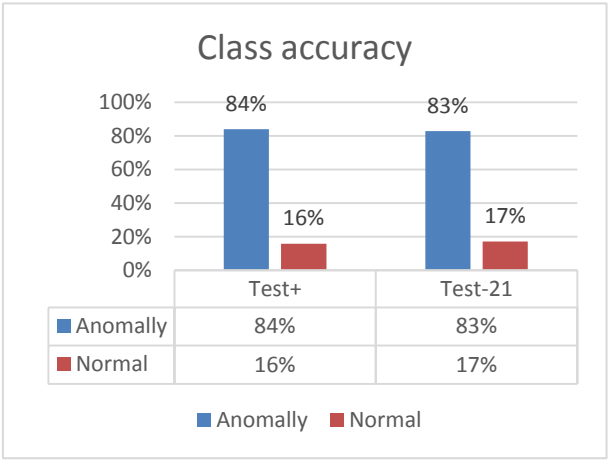


Figure 10: Evaluation (Class Accuracy)

The evaluation of results base on class of accuracy from the study using NSL-KDD Dataset test + and test+21 shows that our developed model shows a low classification of 16% for normal packet and 84% classification for anomalies as shown in figure 10 . The model was able to identify anomalies with 84% Accuracy which is very good considering the focus on detecting suspicious packet.

5. CONCLUSION

Intrusion Detection is based on the premise that the network behavior of attackers or intruders in a network would be different from the behavior of legitimate users of the network. Hence detecting and studying the patterns in these behaviors could very well indicate the presence of an intruder in a network. The developed model was able to classify normal packets and detect anomalies of packets. The goal is to classify incoming network packets to determine whether it is a normal packets or an anomaly. This would improve the security of the network Model.

6. Acknowledgment

The authors wish to thank the Department of Computer Science, University of Ibadan, Nigeria for the support in this research work

REFERENCES

[1] P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges", *Computers & Security*, vol. 28, pp. 18-28, Mar. 2009.

[2] C. Tsai, Y. Hsu, C. Lin, W. Lin "Intrusion detection by machine learning: A review" *Expert Systems with Applications* Volume 36, Issue 10, Pages 11994-12000, Dec. 2009.

[3] Z. Alom, T. Taha "*Network Intrusion Detection for Cyber Security using Unsupervised Deep Learning Approaches". IEEE National Aerospace and Electronics Conference*. 63–69. Jun. 2017.

[4] A. Javaid, Q. Niyaz, W. Sun, & M. Alam, "A deep learning approach for network intrusion detection system". In *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies* pp. 21-26, May. 2016

[5] J. Kevric, S. Jukic, A. Subasi "An effective combining classifier approach using tree algorithms for network intrusion detection." *Neural Computing and Applications*, vol *28*, pp. 1051–1058, Jun. 2017.

[6] Z. Alom, V. Bontupalli, V, T. Taha "Intrusion detection using deep belief networks" *IEEE National Aerospace and Electronics Conference,* pp. 339-344, Jun. 2015.

[7] D. Eduardo, D. Emiro, A. Ortiz, J. Ortega, B. Prieto "PCA filtering and probabilistic SOM for network intrusion detection." *Neurocomputing*, vol. *164*, 71–81, Sep. 2015

[8] D. Eduardo, D. Emiro, A. Ortiz, J. Ortega, B. Prieto "Deep learning approach for Network Intrusion Detection in Software Defined Networking." *Proceedings - International Conference on Wireless Networks and Mobile Communications, WINCOM 2016: Green Communications and Networking*, pp 258–263, Oct. 2015

[9] K. Kato, V. Klyuev "Development of a network intrusion detection system using Apache Hadoop and Spark." *IEEE Conference on Dependable and Secure Computing*, pp 416–423, Aug. 2017

[10] J.Ukam, O. Adeniji, "Performance Evaluation of Error Rate in Immune Inspired Concepts with Neural Network for Intrusion Detection in Cybersecurity" . *IJARCCE*, v 9, pp 16-22, Jun.2020.

[11] Olushola D Adeniji, Olubukola Adigun, Omowumi O Adeyemo, An intelligent spam-scammer filter mechanism using bayesian techniques. International Journal of Computer Science and Information Security (IJCSIS), Vol. 10, No. 3 pp 126, 2013.

[12] K. B. Logunleko, O. D. Adeniji, A. M. Logunleko, .A Comparative Study of Symmetric Cryptography Mechanism on DES, AES and EB64 for Information Security. International Journal of Scientific Research in Computer Science and Engineering Vol.8, Issue.1, pp.45-51, 2020.

[13] O. D. Adeniji, O. O. Olatunji. Zero Day Attack Prediction with Parameter Setting Using Bi Direction Recurrent Neural Network in Cyber Security. International Journal of Computer Science and Information Security (IJCSIS), Vol. 18, No. 3,pp 111-118, 2020.

[14] S.D. Adeniji, S. Khatun, M. A. Borhan, R. S. A. Raja, A design proposer on policy framework in IPV6 network. IEEE International Symposium on Information Technology. Vol. 4, pp. 1-6, 2008.

[15] O. D. Adeniji, Adenike Osofisan, *Route Optimization in MIPv6 Experimental Test bed for Network Mobility: Tradeoff Analysis and Evaluation***.** International Journal of Computer Science and Information Security (IJCSIS), Vol. 18, No. 5, pp 19-28, 2020.

[16] A. A. Olabisi O. D. Adeniji, Abeng Enangha " A Comparative Analysis of Latency, Jitter and Bandwidth of IPv6 Packets Using Flow Labels in Open Flow Switch in Software Defined Network" Afr. J. MIS, Vol.1, Issue 3, pp. 30-36.(2019)