Cyber Security Experts Association of Nigeria (CSEAN)
Society for Multidisciplinary & Advanced Research Techniques (SMART)
Faculty of Computational Sciences & Informatics - Academic City University College, Accra, Ghana
SMART Scientific Projects & Research Consortium (SMART SPaRC)
Sekinah-Hope Foundation for Female STEM Education
ICT University Foundations USA

Proceedings of the Cyber Secure Nigeria Conference – 2023

Performance Evaluation of Random Forest Algorithm for DDoS Attacks Detection in Wireless Sensor Network Using Numericalization and Normalization as Feature Engineering Techniques

[1]Sulaiman, H., [2]Oyefolahan, I.O., [3]Bashir, S.A., [4]Kolo, I.M. & [5]Ndunagu J.N.
Department of Computer Science[1,3,4,5]; Department of Information Technology[2]
School of Information and Communication Technology
Federal University of Technology, Minna
[1,2,3,4], National open University of Nigeria, Abuja[5].
E-mails: [1]sulaiman.pg207857@st.futminna.edu.ng, [2]o.ishaq@futminna.edu.ng,
[3]bashirsulaimon@futminna.edu.ng, [4]idris.kolo@futminna.edu.ng, [5]jndunagu@noun.edu.ng
Phone Nos: [1]+2348068231662, [2]+2348164134704, [3]+2349097622911, [4]+2347039780788

ABSTRACT

Distributed Denial-Of-Service (DDoS) attacks in Wireless Sensor Networks (WSNs) aims to decrease the availability of a service by exhausting the network resources available for traffic, thus preventing legitimate users from accessing the services. This paper evaluates the performance of Random Forest algorithm for distributed denial of service attacks detection in Wireless Sensor Networks (WSNs) with numericalization and normalization as feature engineering techniques and exploratory data analysis to build the Distributed Denial of Service (DDoS) attacks detection model. The dataset was streamlined for optimum performance of the selected algorithms. The experimental result shows the Random Forest achieved an accuracy of 90%, F1 score of 90%, precision of 91% and recall of 90%. The outcomes revealed that the random forest model performed better than the competing models in terms of accuracy, F1 score, precision, and recall. The results of this study shed light on the efficiency of machine learning models in identifying DDoS attacks and suggest areas for future study. The study stresses the significance of data preparation and cleaning in the process and offers a framework for the application of machine learning models in network intrusion detection systems.

Keywords: Wireless Sensor Networks, Distributed Denial of Service, Intrusion Detection System, Random Forest, Machine Learning.

Proceedings Citation Format
Sulaiman, H, Oyefolahan, I.O., Bashir, S.A., Kolo, I.M. & Ndunagu J.N. (2023): Performance Evaluation of Random Forest Algorithm for DDoS Attacks Detection in Wireless Sensor Network Using Numericalization and Normalization as Feature Engineering Techniques. Proceedings of the Cyber Secure Nigeria Conference. Nigerian Army Resource Centre (NARC) Abuja, Nigeria. 11-12th July, 2023. Pp 19-26. https://cybersecurenigeria.org/conference-proceedings/volume-2-2023/
dx.doi.org/10.22624/AIMS/CSEAN-SMART2023P2.

1.. INTRODUCTION

Growing attention is given to Wireless Sensor Network (WSN) due to its wide deployment in military and civilian services. Such services include area monitoring, battlefields, environmental/earth sensing, manufacturing and industrial monitoring, telecommunication, healthcare, smart homes, smart cities, smart transportation and Internet of Things (IoT). Such important services will heighten the interest of security attackers in such networks and necessitate producing continuous security solutions. (Sravanthi & Rama, 2020). Wireless Sensor Networks (WSNs) are cyber-physical systems composed of potentially large quantities (tens to thousands) of sensor nodes that are spatially distributed across large areas and rely on wireless communication. The main purpose of the majority of WSNs is environmental sensing and monitoring. (Lama & Saad, 2021).

Intrusion detection is an area of cyber security dealing with security intrusions, i.e. unwanted breaches of security caused by unauthorized use of resources in computers and computer networks. They also compromise data and communication confidentiality, availability, and integrity (CAI) and can cause a denial of service (Kanimozhi & Prem, 2019). As a result of traditional IDS' poor detection capability and lower accuracy rate, cybercriminals have increasingly used innovative infiltration techniques to circumvent classical IDS mechanisms (Mohammed et al., 2020). Artificial intelligence introduced machine learning algorithms as a technique that provides huge adaptability benefits in wireless sensor networks, Artificial intelligence techniques are performing better accuracy than the traditional methods to detect intrusion for various attacks. The performance of AI techniques-based detection systems for DDoS attacks in WSNs is remarkable (Mohammed et al., 2020).

Machine Learning (ML) is the sub-area of AI. ML refers to the problems it will provide solutions to by recognizing patterns from the database. To collect the input data ML create the algorithms and use statistical analysis for the prediction of outputs with the occurrence of new data (Nagamallik et al., 2020). Machine learning (ML) techniques are considered to be one of the prominent methods that could be used with IDSs to improve their ability to identify and recognize attackers. The ML classification method has been used for DoS detection in WSNs. (Thanh & Vijay, 2019).

2. RELATED WORK

Emmanuel and Olawale (2020) worked on an Improved Genetically Optimized Neural Network Algorithm for the Classification of Distributed Denial of Service Attack; in this research the experiments show the optimization of an improved genetic algorithm with a neural network for the classification of the DDoS has better performance in terms of accuracy and false alarm rate. To implement and measure the performance of this research, the standard KDD99 benchmark dataset was obtained and trained with a new model. The model gives reasonable accuracy and false alarm rate preferable to the existing ones.

The results of this research show new model NN-GA has better accuracy of 98.58 with a lower false positive rate of 0.351 as against the conventional neural network which yielded an accuracy of 88.50 with a false alarm rate of 0.558. The research lacks an improved fitness function or heuristic ensemble with potential to give better detection and false alarm rate, future work needs to improve the model on that. While Nada and Dina (2021) proposed a model of multilayer machine learning-based Intrusion Detection System for Wireless Sensor Networks, stated that intrusion detection in wireless sensor networks is a very challenging task. The majority of the current WSN intrusion detection models were using machine learning methods, but they apply only one method for the whole network. In this paper, a multi-layer framework for intrusion detection systems in WSN was used.

The proposed model consists of two consequent protection layers; the first layer is located on the edge of the network where the sensors are located. It used the Naive Bayes classifier where the traffic was classified into normal or malicious traffic which achieves simplicity and time efficiency in the decision-making process. While the second layer was located on the cloud, and mainly handles suspicious traffic by using a multi-class Random Forest classifier. The weakness of the model indicated in the case of the True Negative Rate (TNR) of the Normal packets where the model achieved a slightly lower value. Feature work will improve the performance of the multilayer detection model in WSN by using one of the deep learning techniques in the second layer where the higher number of attacks types appears.

Xiaopeng et al. (2019) proposed a model on wireless sensor networks intrusion detection based on smote and random forest algorithm. Due to the class imbalance in KDD Cup 99 dataset, this study combines the SMOTE with the random forest algorithm and proposes an ensemble classifier for imbalanced datasets. Experiments on KDD Cup 99 dataset show that the classification accuracy of the random forest algorithm has reached 92.39%, which is higher than other classification methods, such as J48, LibSVM, Naïve Bayes, Bagging, and AdaboostM1. After combining with the SMOTE, the classification accuracy of the random forest has increased to 92.57%, which improves the classification effect of minority classes. The random forest method combined with the SMOTE provides an effective solution to solve the problem of class imbalance and improves the classification accuracy of intrusion detection. Moreover, this method is simple to implement and has strong generalization ability.

It can be widely used in the field of the security of wireless sensor networks to improve the effect of intrusion detection for wireless sensor networks. This research lacks new classification methods. Future work needs to further improve the recognition effect of the intrusion data of wireless sensor networks. Meanwhile Suleiman et al. (2019) worked on the performance of the support vector machine, has been optimized using Cat Swarm Optimization Algorithm. The NSL-KDD dataset was used. The entropy value of each of the attributes was calculated concerning the class value. Attributes with insignificant entropy values were removed during the preprocessing stage. The classification was done with the optimized SVM-CSO. The classification result shows that the CSO-SVM has better performance in all areas compared to the performance of the baseline classifier (Zero R). In terms of accuracy, and F-measure the CSOSVM performs better compare to other clarification algorithms like the popular J48, Naïve Bayes and Random Tree.

Most importantly, the CSO-SVM has a low false positive rate of 0.02 compared to IG-PSO-SVM and IG-ABC-SVM with 0.04 and 0.03 respectively. The limitation is that performance of the model was evaluated with only SVM algorithms. Future work will improve performance of the model with other classification algorithms.

3. METHODOLOGY

This paper used Random Forest Tree classification algorithm to build the Distributed Denial of Service attack (DDoS) classification model. Random forest is a supervised Machine Learning method that is commonly used in regression and classification applications and delivers excellent results most of the time even without hyper parameter modification. It creates several decision trees on various samples and picks the majority vote (Saini, 2021). During training, random forests (RF) build numerous independent decision trees. The final prediction is made by pooling predictions from all trees; the mode of the classes for classification or the mean prediction for regression.

The more significant the characteristic the higher the value. Scikit-learn machine library determines a node's significance using Gini Importance for each decision tree, assuming only two child nodes (binary tree):

$$Ni_j = W_jC_j - W_{left(j)}C_{left} - W_{right(j)}C_{right(j)} \quad (i)$$

Where $ni_{(j)}$ represents the importance of node j, $w_{(j)}$ is the weighted number of samples reaching node j, $C_{(j)}$ is the impurity value of node j, $Left_{(j)}$ represents the child node from left split on node j and $Right_{(j)}$ is the child node from right split on node j.

4. DATA COLLECTION

The data collection phase started by collecting the required NSL-KDD datasets, an updated version of the KDD CUP 99 dataset, which the University of New Brunswick creates to evaluate Machine Learning-based IDSs. This dataset comprises standard forms of attacks (DDoS, DoS, Probing, U2R, and R2L). Before commencing the pre-processing step, the data collected from the mentioned source is converted from an Attribute-Relation File Format (ARFF) to Comma Separated Value (CSV) format to make the data analysis easier in Python.

5. NUMERICALIZATION/VECTORIZATION

Numericalization is the process of converting non-numeric data into numbers. In machine learning, the process of turning textual data into numerical data is known as vectorization or numericalization. It is an essential task since machine learning techniques cannot be used directly on text because they only handle numerical input. This process involves converting the text data into numbers for classification purposes.

5.1. Normalization

Normalization is a method that is frequently used in data preparation for machine learning. The purpose of normalization is to convert the values of numeric columns in the dataset to a similar scale without distorting the ranges of values or losing information. The data in the dataset is normalized using this approach before Training.

6. PSEUDO CODE FOR THE MODEL

The steps or algorithm used to describe the Random Forest-based DDOS detection model is depicted in the pseudo-code below.

i. Algorithm: Random forest algorithm for DDOS attack detection.
ii. Input: Portmap1.csv, LDAP11.csv, and UDPLag11.csv dataset [Training Dataset, Testing Dataset], csv format
iii. Output: Accuracy, recall, precision and F1-score of the trained Random Forest model on test dataset
iv. Apply feature engineering using seaborn, matplotlib,pandad and numpy.
v. Let T0 denote the feature set
vi. For x in dataset do:
vii. Let Tx be the feature set of sample x
viii. For k in x do:
ix. Vk Compute (k,w)
x. Append Vk to Tx
xi. Append Tx to T0
xii. Apply cross-validation Split feature set into Xtrain , Xtest , Ytrain, and Ytest
xiii. Model = RandomForest(Xtrain, Ytrain)
xiv. Confusion Matrix = (p, Ytest, Model)
xv. Compute Accuracy, recall, precision, and F1-score
xvi. Output Performance score

7. DISCUSSIONS OF RESULTS

Figure 2 below generates a bar graph that displays the relative weights of several dataset attributes. It shows the top 20 characteristics, as identified by the Extra Trees Classifier mathematical model. The size of the bars on the graph denotes the significance of each variable in predicting a certain result. This graph was produced using a standardized dataset, which implies that the data has been normalized to facilitate feature comparisons. This graph serves as a decision-making tool for future research by aiding in the identification of the dataset's most important elements.

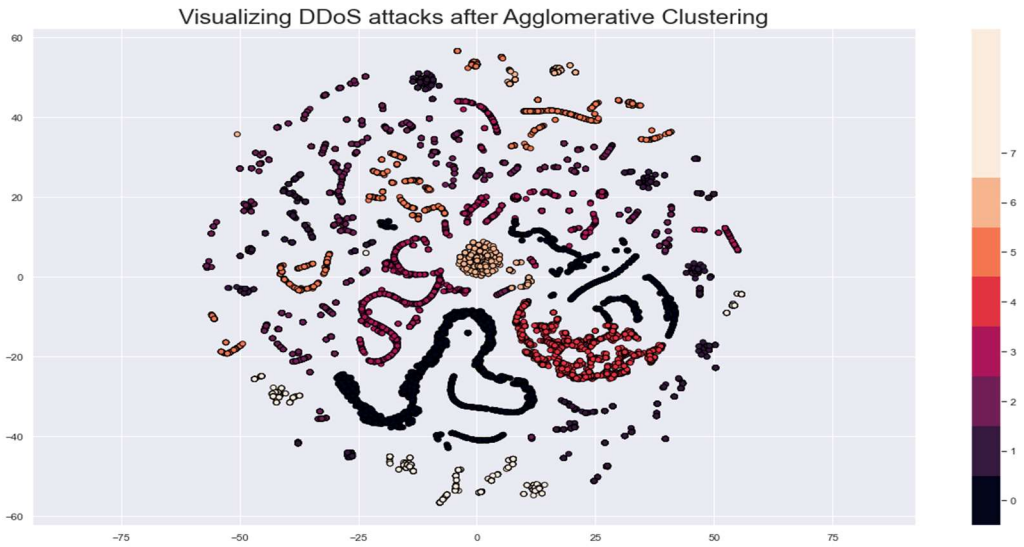

Figure 2 DDOS attack Visualization using Agglomerative Clustering.

8. RANDOM FOREST (RF)

This section describes the performance of the Random Forest Model. The confusion matrix is a performance measurement for our classification problem. It shows different combinations of predicted and actual values.

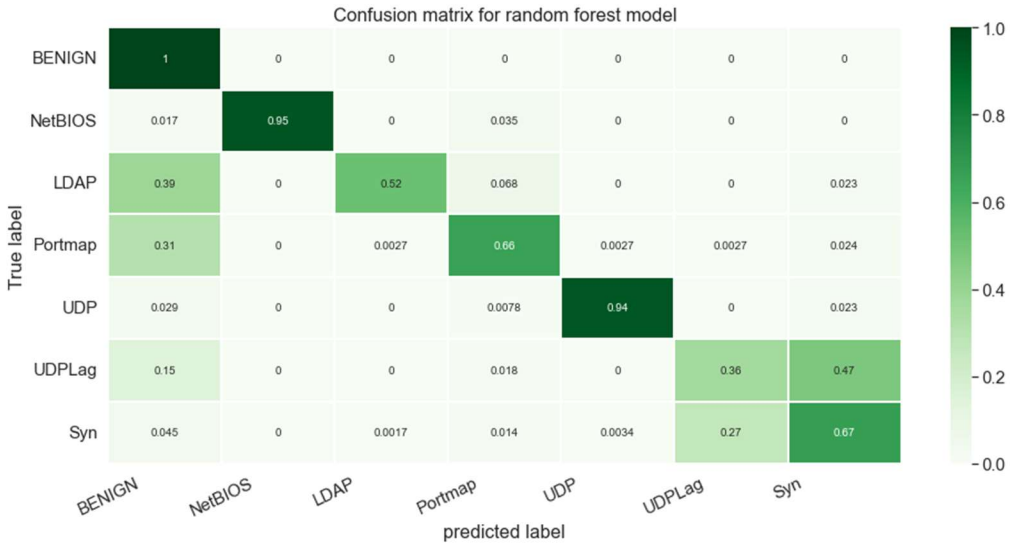

Figure 3 Random Forest Confusion Matrix

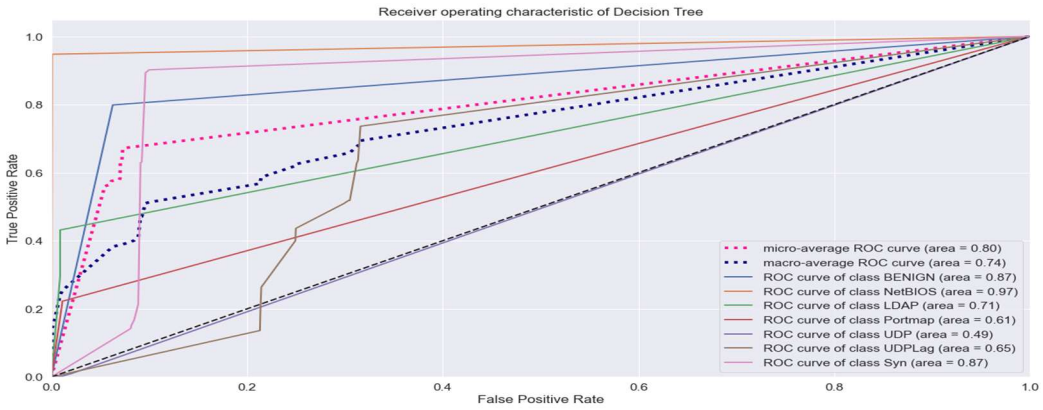

Figure 4 Random Forest ROC

Random Forest achieved an accuracy of 90%, F1 score of 90%, precision of 91%, and recall of 90%.

9. CONCLUSION AND FUTURE WORK

This paper evaluates the Algorithm performance on distributed denial of service attacks in Wireless Sensor Networks (WSNs) using Random Forest Tree classification algorithm to build the Distributed Denial of Service attack (DDoS) classification model. The dataset was streamlined for optimum performance of the selected algorithms. The experimental result shows the Random Forest achieved an accuracy of 90%, F1 score of 90%, precision of 91% and recall of 90%.

The outcomes revealed that the random forest model is the suitable models in terms of accuracy, F1 score, precision, and recall. Based on our review above, there are methods with better performance than these outputs but this research used Numericalization and Normalization as Feature Engineering Techniques; The result of this study shed light on the efficiency of machine learning models in identifying DDoS attacks and offers a framework for the application of machine learning models in network intrusion detection systems. The study recommended that the presented model is subjected to further research.

10. DIRECTION FOR FUTURE WORK

For future work, Focus can be drawn towards hybridizing deep learning and machine learning models to tackle the problems associated with DDoS attack Detection in WSNs. The model presented in this study is subject to further research and can be improved upon to create a better model.

REFERENCES

1. Emmanuel, H. G., & Olawale, S. A. (2020). Improved Genetically Optimized Neural Network Algorithm for Classification of Distributed Denial of Service Attack. *LAUTECH Journal of Computing and Informatics (LAUJCI*) ,1(1), 58-75. www.laujci.lautech.edu.ng
2. Kanimozhi, V., & Prem, J. T. (2019, April 24). Artificial Intelligence based Network Intrusion Detection with hyper-parameter optimization tuning on the realistic cyber dataset CSE-CIC-IDS2018 using cloud computing. ICT Express 5 (2019) 211–214.
3. Lama, A., & Saad, A.-A. (2021, March). Performance Evaluation of Machine Learning Techniques For Dos Detection In Wireless Sensor Network. *International Journal of Network Security & Its Applications (IJNSA)*, 13(2), 21-23.
4. Mohammed, A.-N., Mohammed, A. R., Adamu, A. I., & Hafizur Rahman, M. M. (2020). AI-Based Techniques for DDoS Attack Detection in WSN: A Systematic Literature Review. *Journal of Computer Science*, 16 (6): 848.855 DOI: 10.3844/jcssp.2020.848.855.
5. Nada, M. A., & Dina, M. I. (2021). A Multi-layer Machine Learning-based Intrusion Detection System for Wireless Sensor Networks. *International Journal of Advanced Computer Science and Applications (IJACSA),* 12(4), 281-288.
6. Nagamallik, S. R., Thirupath, N. R., Venkata, N. M., & Debnath, B. (2020). Machine Learning Algorithms to Enhance Security In Wireless Network. *Journal of Critical Reviews,* 7(14), 426-432.
7. Saini, A. (2021). Random Forest Algorithm for Absolute Beginners in Data Science. Retrieved January 6, 2023, from Analytics Vidhyawebsite: https://www.analyticsvidhya.com/blog/2021/10/an- introduction-to-random-forest-algorithm-for beginners/#:~:text=Random%20forests%20are%20a%20supervised
8. Sravanthi, G., & Rama, P. V. (2020). A Study on Intrusion Detection System in Wireless Sensor Networks. International Journal of Communication Networks and Information Security (IJCNIS) Vol. 12, No. 1, April 2020, 127-141.
9. Suleiman, I., Oyefolahan, I. O., & Ndunagu, J. N. October 14-17 (2019). Intrusion Detection Based on Support Vector Machine Optimized with Cat Swarm Algorithm. Proceedings of the 2nd International Conference of the IEEE Nigeria Computer Chapter: IEEEnigComputConf'19: Ahmadu Bello University, Zaria, Nigeria, October 14-17, 2019, 138-145.
10. Xiaopeng, T. S., Zhiping, H., Xiaojun, G., Zhen, Z., Xiaoyong, S., & Longqing, L. (2019). Wireless Sensor Networks Intrusion Detection Based on SMOTE and the Random Forest Algorithm. College of Artificial Intelligence, National University of Defense Technology, Changsha 410073, China; tanxiaopeng14@nudt.edu.cn (X.T.); susj-5@163.com (S.S.); kdhuangzp@163.com (Z.H.); Sensors 2019, 19, 203; doi:10.3390/s19010203.www.mdpi.com/journal/sen, 1-15.