A Model to Detect Network Intrusion Using Machine Learning

J. Palimote[1], L. Atu[2] & E. Osuigbo[3]

[123]Department of Computer Science

Kenule Beeson Saro-wiwa Polytechnic, Bori,

River State, Nigeria.

Abstract-*The attacks on computer security are becoming global and it is a very important security threats issues to the cyberspace, that if an organization is not mindful of, important data will be accessed, modified, or deleted. Computer security attackers utilize the compulsions and security weaknesses in the computer network and data system to carry out an attack, which ideals the divulgence of system information and the capture of the privacy of users and threaten data availability or integrity. The proposed system aim in developing a model to detect network intrusion using machine learning algorithm. The dataset consists of different categories of intrusions stimulated in a military environment. The dataset consists of a raw TCP/IP data for a network by stimulating a typical US Air Force Lan. The dataset is made up of 41 Columns. The class column consists of two types, which are Normal Network Packets and Anomalous Network Packets. The dataset was preprocessed by converting some features that was characters as values to 0s and 1s by using the pandas.get_dummies function. After that, we applied feature extraction by dropping and adding some features. The number of columns increased to 112 columns after performing feature extraction. The dataset was split into a train and a testing set using the train_test_split function. We use two machine learning algorithms in building/training our model. These machine learning algorithms are Random Forest Classifier and Support Vector Machine. After training/building our models, Random forest Classifier had the highest accuracy results which is about 99.78% while Support Vector Classifier had about 96.87%.*

Keywords- Network Intrusion, Machine Learning, Support Vector Classifier, Random Forest Classifier.

I. Introduction

The attacks on computer security are becoming global and it is a very important security threats issues to the cyberspace, that if an organization is not mindful of, important data will be accessed, modified, or deleted. Computer securityattackers utilize the compulsions and security weaknesses in the computer network and data system to carry out an attack, which ideals the divulgence of system information and the capture of the privacy of users and threaten data availability or integrity [1]. Computer security attacks is still growing fast, picking out system data, networks system, personal end-device, and commercial infrastructures. Intrusion detection can be defined as the process to detect the attempted intrusion attacks, in progress intrusion attacks, or breaching.Because of expansions in the volume and refinement of attacks carried out on network systems, constructive methods and instruments are needed to keep away from the dangers of revelation of important data, interruption, and unapproved access. Network intrusion detection system (NIDS) has become fundamental for some associations. Such system monitors the traffic in network and identify unusual exercises or the attacks on computer securities to guarantee the wellbeing and security of their correspondence and data [2].

Intrusion attack is categorized into two types. Which are signature-based intrusion detection system and Anomaly based intrusion detection system. These days, the standard protection approaches depends on signature base intrusion detection system, which uses a pattern matching method to investigate and detect threats from the approaching network packet by contrasting a preinstalled attack signature on their database against the approaching network activities. Signature based intrusion detection system has a great advantage in detecting network attacks, with a small false alarm. Nonetheless, it requires a pre-introduced signature and it is unaware of unknown network attacks. When making use of signature based intrusion detection system for network security, regular support and routine updates are required. Interestingly, Anomaly based intrusion detection system centers around deviations of the traffic example and utilize those deviations to assess approaching traffic and decide the opportunity of abnormality, in any event, when confronted with

unknown attacks [3]. To protect the immense network intrusion, scholars have completed a ton of investigation. An intrusion detection system (IDS) is a cyber-security gadget that performs continuous checking of transmissions of network also, issues alarms. In addition, it takes the proactive reaction measures when dubious transmissions are found. The IDS varies from other network security gadgets in that it possessed the forward-looking security protection innovation [4]. Yet, confronted with the violently extreme the internet security circumstance, the conventional intrusion detection approach has steadily uncovered numerous downsides against securing of network security. The normal imperfection is the presences of more genuine False Positive (FP) and False Negative (FN). This paper presents a model to detect network intrusion using machine learning algorithm.

2. Related Works

CANnolo: An Anomaly Detection System based on LSTM Autoencoders for Controller Area Network [5] proposed IDS dependent on Long Short-Term Memory (LSTM) -autoencoders to detect inconsistencies in Controller Area Networks (CANs). During the preparation stage, CANnolo naturally examines the CAN streams and constructs a model of the genuine information successions. At that point, it identifies abnormalities by registering the distinction between the remade and the individual genuine groupings. CANnolo was tentatively assessed on a bunch of reproduced attacks applied over a real world dataset. Their experimental results show that their proposed system beats the best in class model by improving the detection rate and accuracy.

Semantic Models for Network Intrusion Detection [6] discussed the validation and design of the hierarchical intrusion detection system using machine learning techniques, thy carried out a philosophical attack on network system, which allow them to detect and classify the attack types. The trained model was evaluated and compared to similar techniques, which they had 0.998% accuracy.

A Deep Reinforcement Learning Approach for Anomaly Network Intrusion Detection System [7] Presented a deep reinforcement learning-based (DRL) for anomaly network intrusion detection system. Their proposed system has the capability of self-updating to consider new types of traffic network conduct. They made three major contributions on their study. First, they discussed the overall application, of which their method can be applied to; they demonstrated their work using well-known NIDS benchmark datasets: UNSW-NB15 and NSL-KDD, and a real campus network log. Second, they demonstrated the feasibility of their method by comparing their method with three other classic machine learning methods and two related published results which they had an accuracy of about 98.7%. Third, their network model has the ability to process a million scale of network traffic on a real-time basis.

Banking Intrusion Detection Systems based on customer's behavior using Machine Learning algorithms: Comprehensive study [8] developed multi-stage checking system and made use of machine-learning techniques in obtaining the anomaly based Intrusion detection on network system. Applied these stages successively on Ids dataset to detect how the user behaves anomalous. Their proposed system and the experimental results showed the efficiency of the system in the mission of detection, distinguish attack, and define normal behavior that their system had a high detection rate of about 98%.

An Intrusion Detection Method Based on Decision Tree-Recursive Feature Elimination in Ensemble Learning [9] presented an anomaly based network intrusion technique using Decision Tree-Recursive Feature Elimination (DT-RFE) feature in ensemble learning. Firstly, they processed their data using the Decision Tree-Based Recursive Elimination Algorithm in reducing features, selecting appropriate features, removing redundant values and infinite data from the dataset in other to achieve a better training result and performance. They also made use of an ensemble learning called stack, they combined Stack ensemble learning algorithm, Decision Tree (DT) with Recursive Feature Elimination (RFE) techniques. After successfully carrying out series of comparison results by cross-validation on the KDD CUP 99 and NSL-KDD datasets, indicate that the DT-RFE and Stacking-based approach can better improve the performance

of the IDS, and the accuracy for all kinds of features is higher than 99%, except in the case of U2R accuracy, which is 98%.

Intrusion detection using machine learning: A comparison study [10] presented an Intrusion Detection System for network using machine leaning approach, with a good combination of feature extraction method and classifier by lucubrating the union of most of the popular feature extraction methods and algorithms. They made selection on some important features from the original dataset using feature extraction techniques and then the set of important features is used to train different types of algorithms to make the Intrusion Detection System. They carried out a Five folds cross validation on NSL-KDD dataset to find results. It is finally observed that K-Nearest Neighbor algorithm had better performance than others and, among the feature extraction techniques, data gain ratio based feature extraction technique is better.

Network Intrusion Detection Using Machine Learning [11] proposed two machine-learning algorithms to detect any anomalous behavior in found on the network system. The overall performance of the proposed system is distinguished by carrying out an evaluation on the detection accuracy, false negative rate, false positive rate, and time it takes in detecting the intrusion attack. The proposed system showed the productiveness of the classifier in detecting/identifying the intrusion with higher detecting accuracy of 98.76% and low false positive rate of 0.09% and false negative rate of 1.15%, whereas the normal Support Vector Machine based scheme achieved a detection accuracy of 88.03% and false positive rate of 4.2% and false negative rate of 7.77%.

Cyber Intrusion Detection Using Machine Learning Classification Techniques [12] employed different well known machine learning classification algorithms, namely Bayesian Network, Naive Bayes classifier, Random Decision Forest, Decision Tree, Random Tree, Decision Table, and Artificial Neural Network, in detecting network intrusions due to given intelligent services in the domain of computer-security. Finally, they tested the efficiency of different kind experiments conducted on computer-security datasets, having different types of attacks on network system and check the efficiency of the performance metrics, precision, recall, f1-score, and accuracy. The accuracy are Decision Tree (93%), Random Forest, (94%), Naïve Bayes (91%) and Artificial Neural Network (91%).

3. Methodology

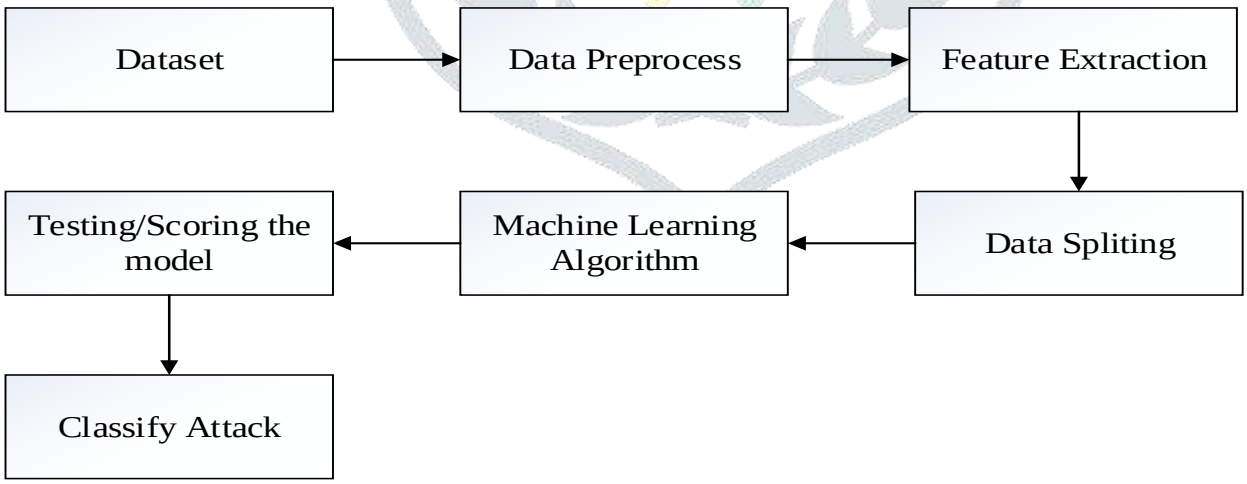

Figure 1: Architecture of the Proposed System Design

The proposed system uses a Network Intrusion dataset, which was downloaded from kaggle.com. The dataset consists of different categories of intrusions stimulated in a military environment. The dataset consists of a raw TCP/IP data for a network by stimulating a typical US Air Force Lan. The dataset is made up of 41 Columns. The class column consists of two types, which are Normal Network Packets and

Anomalous Network Packets. The dataset was preprocess in removing null values, infinite values and also converting some characters to 0s and 1s. Feature extractions was used in bringing all columns to pre-defined functions. The train data will be divided into a training part and a testing part using the train_test_split function from scikit-learn library. After successfully splitting of data, the splitted will be passed to the machine learning algorithm, which are Random Forest Classifier and Support Vector Classifier. The trained model will be scored based on accuracy, precision, false positive, and false negative. The trained model will be tested by entering some network packets and also, by checking the classified network attacks performed by the trained model.

Table 1: Evaluation metric

Definition	Formula
Accuracy	$\frac{TP+TN}{TP+FN+TN+FP}$
Error	1-accuracy
Recall	$\frac{TP}{TP+FN}$
Precision	$\frac{TP}{TP+FP}$
F-measure	$\frac{2\,Precison \times Recall}{Precision+Recall}$

4. Discussion of Result

The proposed system uses a Network Intrusion Dataset, the dataset to be audited was provided which consists of a wide variety of intrusions simulated in a military network environment. It created an environment to acquire raw TCP/IP dump data for a network by simulating a typical US Air Force LAN. The LAN was focused like a real environment and blasted with multiple attacks. A connection is a sequence of TCP packets starting and ending at some time duration between which data flows to and from a source IP address to a target IP address under some well-defined protocol. Also, each connection is labelled as either normal or as an attack with exactly one specific attack type. Each connection record consists of about 100 bytes.

For each TCP/IP connection, 41 quantitative and qualitative features are obtained from normal and attack data (3 qualitative and 38 quantitative features) .The class variable has two categories, which are Normal and Anomalous. The dataset was preprocessed by converting some features that was characters as values to 0s and 1s by using the pandas.get_dummies function. After that, we applied feature extraction by dropping and adding some features. The number of columns increased to 112 columns after performing feature extraction. The dataset was split into a train and a testing set using the train_test_split function. We use two machine learning algorithms in building/training our model. These machine learning algorithms are Random Forest Classifier and Support Vector Machine. After training/building our models, Random forest Classifier had the highest accuracy results which is about 99.78% which can be seen in figure 5 while Support Vector Classifier had about 96.87% which can be seen in figure 6.

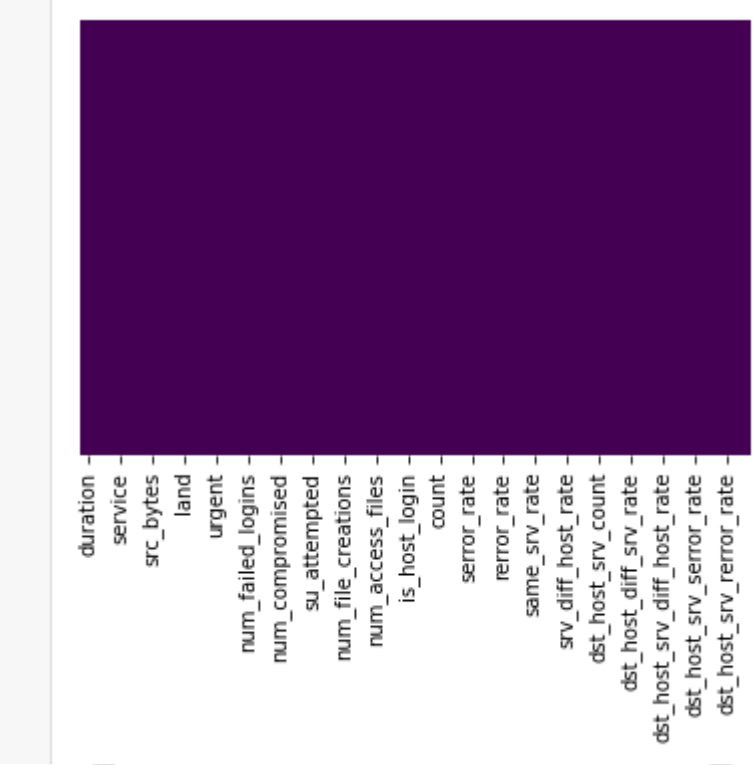

Figure 2:The dataset has been cleaned

The dataset was cleaned by making sure that null values, infinite values are not present in the data.

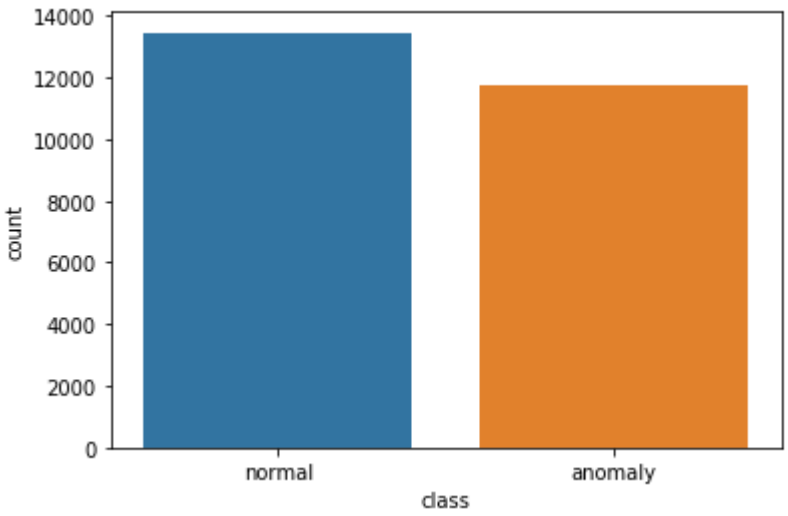

Figure 3: A count plot of the dataset that contains both a normal network packets and an Anomalous Network Packets. The histogram shows that about 13000 network packets were a normal one while about 11000 network packets were anomalous.

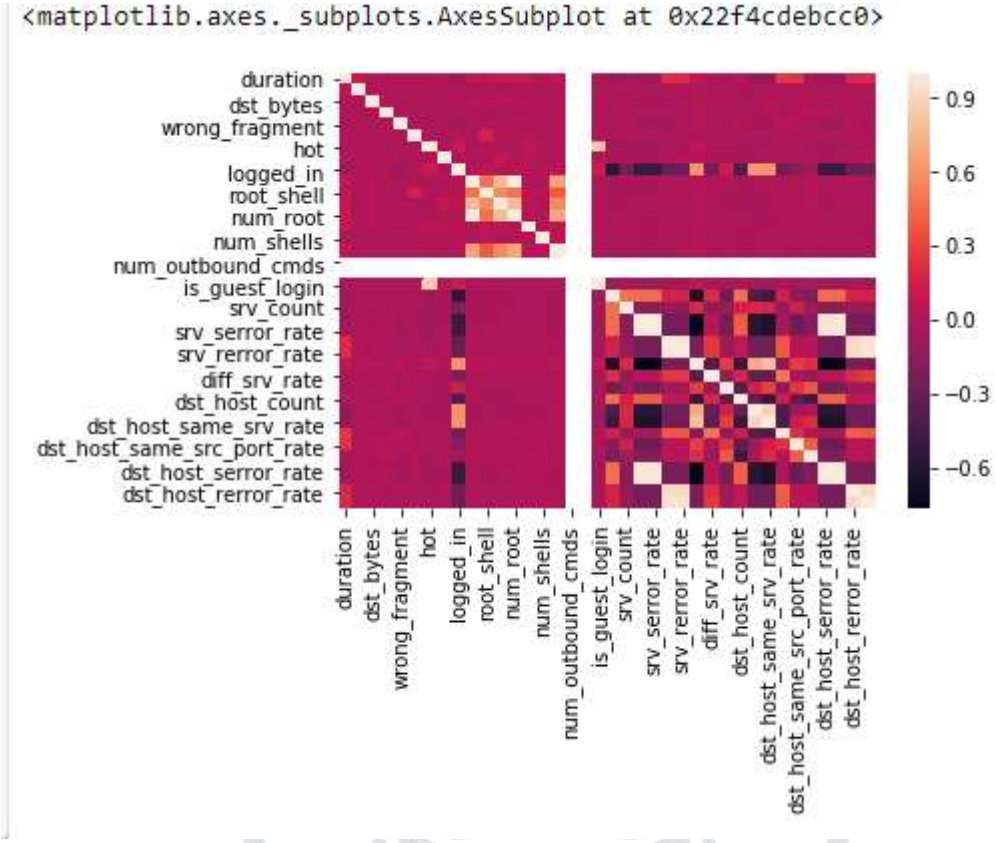

Figure 4: A correlation metrics of the dataset.

	precision	recall	f1-score	support
0	1.00	1.00	1.00	3836
1	1.00	1.00	1.00	4478
accuracy			1.00	8314
macro avg	1.00	1.00	1.00	8314
weighted avg	1.00	1.00	1.00	8314

Figure 5: Metrics evaluation of Random Forest Classifier, which have an accuracy of about 99.78%. This was approximated to be 100% accuracy

	precision	recall	f1-score	support
0	1.00	0.93	0.96	3836
1	0.94	1.00	0.97	4478
accuracy			0.97	8314
macro avg	0.97	0.96	0.97	8314
weighted avg	0.97	0.97	0.97	8314

Figure 6

Metrics evaluation of Support Vector Classifier, which have an accuracy of about 96.7%. This was approximated to be 97% accuracy.

5. Conclusion

The attacks on computer security are becoming global and it is a very important security threats issues to the cyberspace, that if an organization is not mindful of, important data will be accessed, modified, or deleted.The proposed system uses a Network Intrusion dataset, which was downloaded from kaggle.com. The dataset consists of different categories of intrusions stimulated in a military environment. The dataset consists of a raw TCP/IP data for a network by stimulating a typical US Air Force Lan. The dataset is made up of 41 Columns. The class column consists of two types, which are Normal Network Packets and Anomalous Network Packets. The dataset was preprocessed by converting some features that was characters as values to 0s and 1s by using the pandas.get_dummies function. After that, we applied feature extraction by dropping and adding some features. The number of columns increased to 112 columns after performing feature extraction. The dataset was split into a train and a testing set using the train_test_split function. We use two machine learning algorithms in building/training our model. These machine learning algorithms are Random Forest Classifier and Support Vector Machine. After training/building our models, Random forest Classifier had the highest accuracy results which is about 99.78% while Support Vector Classifier had about 96.87%.

References

[1]. D. E. Denning, “An intrusion-detection model,” IEEE Transactions on Software Engineering, vol. 13, no. 2, pp. 222–232, 1987.

[2]. B. Deokar and A. Hazarnis, "Intrusion Detection System using Log Files and Reinforcement Learning," International Journal of Computer Applications, vol. 45, no. 19, 2012.

[3]. K. Sethi, E. S. Rupesh, R. Kumar, P. Bera and Y. V. Madhav, "A contextaware robust intrusion detection system: a reinforcement learning-based approach," International Journal of Information Security, 2019.

[4] L. N. Tidjon, M. Frappier, and A. Mammar, “Intrusion detection systems: a cross-domain overview,” IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3639–3681, 2019.

[5]. S. Longari, D. H. N. Valcarcel, M. Zago, M. Carminati and S. Zanero, "CANnolo: An Anomaly Detection System based on LSTM Autoencoders for Controller Area Network," in IEEE Transactions on Network and Service Management, doi: 10.1109/TNSM.2020.3038991.

[6]. P. Bednar, M. Sarnovsky, P. Halas “Semantic Models for Network Intrusion Detection”, Ist Eclipse Research International Conference on Security, Artificial Intelligence and Modelling for the next generation internet of things, 2020.

[7]. H. Ying-Feng, M. Matsuoka “A Deep Reinforcement Learning Approach for Anomaly Network Intrusion Detection System”, 9th IEE International Conference on Cloud Networking 2020.

[8]. W. S. Mahdi, A. T. Maolood “Banking Intrusion Detection Systems based on customers behavior using Machine Learning algorithms: Comprehensive study” Journal of Al-Qadisiyah for Computer Science and Mathematics Vol.12(4), pp.1–11, 2020.

[9]. W. Lian, G. Nie , B. Jia , D. Shi, Qi Fan, Y. Liang “An Intrusion Detection Method Based on Decision Tree-Recursive Feature Elimination in Ensemble Learning”, Mathematical Problems in Engineering Volume 2020,pp.1-15. 2020.

[10]. S.k. Biswa “Intrusion detection using machine learning: A comparison study”, International Journal of Pure and Applied Mathematics Volume 118 No. 19 2018, 101-114, pp. 101-112, 2018.

[11]. M. N. Chowdhury, K. Ferens “Network Intrusion Detection Using Machine Learning”, Proceedings of the International Conference on Security and Management (SAM); Athens : 30-35. Athens: The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp) 2016.

[12]. H. Alqahtani,, I.H. Sarker, A. Kalim, S. Hossain, h. Ikhlaq, S. Hossain “Cyber Intrusion Detection Using Machine Learning Classification Techniques”, International Conference on Computing Science, Communication and Security COMS2 2020: Computing Science, Communication and Security pp 121-131.