Cyber Attack Detection in a Global Network Using Machine Learning Approach

[1]Nureni A. Azeez, [1]Odeyemi O. Taiwo, [1]Isiekwene C. Chioma and [2]Ademola P. Abidoye
[1]Department of Computer Sciences, University of Lagos, Akoka, Nigeria
[2]Kennesaw State University, Atlanta, USA
nurayhn1@gmail.com | taiwo.odeyemi@outlook.com | {isiekwenechioma | ademola.abidoye}@gmail.com

Received: 27-SEP-2023; Reviewed: 25-NOV-2023; Accepted: 01-DEC-2023
http://doi.org/10.46792/fuoyejet.v8i4.1113

ORIGINAL RESEARCH

Abstract- In this digital age, inter-device communication is key to seamless and smooth handshaking. Communication can range from Internet of Things communication (IoT), autonomous vehicles, mobile communication and a plethora of other uses. These communications need to be protected against attacks. Unfortunately, with the widespread use of the internet, cyberattacks have become rampant. This research introduces the use of seven (7) machine- learning models alongside four different ensemble methods to compare the effectiveness of different Machine learning algorithms and ensemble models for intrusion detection. The network traffic was categorized as The Onion Router (TOR or non-TOR) traffic and further categorized if the network traffic data was Benign or Bot/Infiltration traffic data. This was achieved using: – Naïve Bayes, Decision Tree, K-Nearest Neighbor, Logistic Regression, Neural Network, Quadratic Discriminant Analysis, and Support Vector Machine. The ensemble models used are Adaboost, Gradient Boosting, Random Forest, and Max Voting. The "CIC IDS 2017", ("CSE-CIC-IDS2018"), "01-03-2018" and "02-03-2018" datasets were used. For dataset 1, among the regular machine learning models, Decision Trees had the highest values. Accuracy was 97.46% and precision was 89.88%. The highest ensemble performer was the Random Forest ensemble, which had an accuracy of 98.28% and a precision score of 93.20%. For dataset 2, Decision Trees also had the highest accuracy score of 99.86% and a precision score of 99.66%. The highest ensemble performer was the Random Forest ensemble which had an accuracy score of 99.89% and a precision score of 99.70%. For dataset 3, amongst the regular machine learning models, Neural Network had the highest accuracy score of 78.68% and a precision value of 72.92% while the highest ensemble performer was Gradient Boosting with an accuracy of 79.16% and a precision score of 81.25%. The results were shown using line charts and a confusion matrix. From the experiment, it is evident that amongst the traditional Machine Learning Models, Decision Tree- is (or Trees are) the most efficient while the ensemble Models revealed Random Forest as the most efficient of the ensemble models.

Keywords- Cyberattack, Classification, Ensemble Method, Intrusion Detection, Intrusion Prevention, Network.

1 INTRODUCTION

Cyber threats have been one of the top risks as a result of the enormous rise in the quantity and severity of cyber-attacks in recent years (Warraich and Morsi, 2023). An example is the cyber-attack that affected the Democratic National Committee (DNC), where attackers released 19,000 emails and 8,000 attachments. Also, according to Chandrasekar *et al.* (2017), ransomware discovered in 2016 has tripled. Detecting malware by an intrusion detection system can be tricky due to the various tactics employed by its creators. The market for underground services is developing at a faster rate, offering malicious software to criminals. It has developed into a powerful ecosystem designed to take advantage of every opening and flaw in a world that is becoming more interconnected. For instance, malware developers attempted to mine bitcoins by either directly stealing the login information for customers' crypto currency wallets or by using their computing resources (Cleary *et al.,* 2018). The primary function of an intrusion detection system is to distinguish between normal and abnormal network activity while reducing misclassifications (Mohammadpour et al, 2015).

*Corresponding Author

Section B- ELECTRICAL/ COMPUTER ENGINEERING & RELATED SCIENCES
Can be cited as:
Azeez N. A., Odeyemi O. T., Isiekwene C. C., and Abidoye A. P. (2023). Cyber Attack Detection in A Global Network Using Machine Learning Approach, FUOYE Journal of Engineering and Technology (FUOYEJET), 8(4), 448-455
http://doi.org/10.46792/fuoyejet.v8i4.1113

It is therefore imperative to develop an even more effective Intrusion detection system to efficiently detect and prevent cyberattacks. Machine learning has been utilized to improve the operations of intrusion detection systems during these past years. Machine Learning-based systems for identifying cyber threats have significantly improved with the development of artificial intelligence (AI) tools, and they have produced notable outcomes in numerous studies. However, protecting IT systems from threats and criminal conduct is still very difficult due to the continually evolving nature of cyberattacks (Golchha *et. al.,* 2023).

Machine learning approaches are effective in identifying and preventing cyberattacks due to their ability to make an inference from previously available data. These approaches rely on learning the attack model from past threat information and using trained models to find previously unidentified cyber threats (Shen *et. al.,* 2018). Convolutional neural network architectures were tested by (Gibert *et. al.,* 2018) and (Khan *et. al.,* 2018) to identify certain features and patterns that may be utilized to classify malware into various distinct groups. Several other works have been done to detect threats in various kind of environments using machine learning techniques. With the available results regarding the solution to these cyber threats, the solution remains elusive as there are inconsistencies in the current solutions while some are ineffective as proposed by some authors.

2 BACKGROUND

2.1 INTRUSION DETECTION SYSTEM (IDS)

Intrusion is any attempt to compromise "confidentiality,", "integrity," and "availability". Hubballi and

Suryanarayanan (2014) describe an intrusion detection system as one that monitors the activities on a given system and reports violations. An intrusion detection system is software and/or hardware created to identify unauthorized attempts to gain access to, manipulate, or disable a computer system, typically through a network like the Internet. An IDS can do a thorough security analysis, detect and stop malicious attacks on the network, and keep performance normal throughout any outbreak (Abirami and Palanikumar, 2023).

2.2 Classification of Intrusion Detection Systems

Intrusion detection methods are typically divided into three groups:

Signature/Misuse intrusion detection (SIDS): This is based on finding similar characteristics between entities and associating them with the characteristics of a known attack (Khraisat et al., 2018). Signature- or misuse-based intrusion detection is quite adept at identifying existing attacks, but it is essentially ineffective at identifying new attacks whose patterns haven't been recorded in the database yet. To compare an attack pattern with known attack patterns in the database, it uses pattern matching. Matching learning techniques are used in SIDS to identify a prior intrusion. This triggers an alarm when a previously known intrusion is detected (Azeez *et. al.*, 2022).

Anomaly-based intrusion detection (AIDS): Analyzing anomalous behaviour in contrast to regular behaviour allows an anomaly-based intrusion detection system to identify new threats (Husnoo, *et. al*., 2023). It detects new attacks with a comparatively high rate, although numerous false positives are generated.

Hybrid Intrusion detection (HIDS): By having a high detection rate for previous intrusions and the capability of anomaly detectors to identify new attacks, hybrid detection can benefit from misuse detection.

2.3 Categories of Intrusion Detection Systems

There are various kinds of Intrusion Detection Systems namely:

Network intrusion detection systems (NIDS): The local system analyzes incoming network traffic and compares it to a database of known vulnerabilities and attacks. Without interfering with network traffic, NIDS monitors network attacks. Attack signatures are used by NIDS to define the parameters of an attack; network-based systems are free to create their own signatures (Azeez *et. al.*, 2022). Usually, these guidelines or signatures are learned from earlier attacks. Network IDS have the benefits of being simple to set up, inexpensive, and capable of detecting network-based attacks, intrusions, and failed attempts to attack a system (Mukherjee, 2023).

Host-based intrusion detection systems (HIDS): They operate on all the devices connected to the network. It keeps track of file modifications and compares them to earlier snapshots. Additionally, it examines modified files, system settings, and kernel logs. System logging and other data are used to detect intrusions, as are system logs produced by operating system activities. Host-based systems heavily rely on audit trails. Their handlers are referred to as sensors. Higher levels of abstraction make the majority of attacks invisible, but the information above enables the IDS to identify subtle patterns of misuse. Host-based IDS has several advantages over network-based IDS, including the ability to detect and respond to attacks in realtime without the need for additional hardware, the ability to monitor system activities, provide statistics on attacks, and determine whether an attack was successful or not (Azeez *et. al.*, 2021).

Application Protocol Intrusion Detection System (APIDS): APIDS identifies the intrusion by examining the protocol's behavior and event history. A system or agent is inserted between the running processes and the group of servers, analyzing and keeping track of the application protocol used by the connected devices. It monitors communications between applications on a protocol level (Azeez *et. al.*, 2022).

Protocol Based Intrusion Detection System (PIDS): This basically operates on the front end of a server monitoring the communication between devices and the server. It ensures that the security protocol is as specified so as to protect the web server.

Hybrid Intrusion Detection System: This combines various types of Intrusion detection systems to achieve a more effective one.

3 Methodology

3.1 Cyber-attack Detection in a Global Network Using Machine Learning Approach

As earlier stated, traditional Intrusion Detection techniques have hardly been sufficient for Intrusion detection in recent times. Attackers have constantly changed their methods to evade detection. In reality, it is almost impossible to constantly tweak IDS software to handle these changing attack vectors. IDSs which employ the use of Artificial Intelligence are able to predict these changes within record time due to the patterns that have been learned or behaviours gleaned over time which they can then be used to make the required changes or steps to both detect and prevent the attack(s). Machine Learning based IDS are able to carry out this tasks expertly due to the vast amount of data they have been trained with (Azeez and Odejinmi, 2023).

3.2 The Ensemble Learning Approach

This research introduces the use of seven (7) machine-learning models alongside four different ensemble methods to compare the effectiveness of different Machine learning algorithms and ensemble models for intrusion detection. The network traffic was categorized as (TOR or non-TOR) traffic and further categorized if the network traffic data was Benign or Bot/Infiltration traffic data. This was achieved using by using seven (7) machine learning algorithms: – Naïve Bayes, Decision Tree, K-Nearest Neighbor, Logistic Regression, Neural Network, Quadratic Discriminant Analysis, and Support Vector Machine. The ensemble models used are Adaboost,

http://dx.doi.org/10.46792/fuoyejet.v8i4.1113 http://journal.engineering.fuoye.edu.ng/

Gradient Boosting, Random Forest, and Max Voting (Azeez *et. al.*, 2022).

Ordinarily, the traditional machine learning models if and when used individually for building Intrusion detection models are able to detect Intrusions to a great level of accuracy and precision. However, employing the use of Ensemble Learning Models helps to further improve the precision and accuracy of detection. In order to outperform the individual models, an ensemble model mixes different machine learning classifiers. Each "component classifier" is trained to make predictions using the dataset. The result of integrating these forecasts is the final prediction. This conclusion can be reached in a number of ways, including stacking, voting, bagging, and boosting.

3.3 Data Collection

3.3.1 The Canadian Institute for Cybersecurity (CICIDS 2017) Dataset

The "CIC IDS2017" dataset was developed by the Faculty of Computer Science, University of New Brunswick, in 2017. It is a modified version of the ISCX 2012 dataset (Hossein Hadian Jazi, et al 2017; University of New Brunswick, 2012). The CIC IDS2017 dataset fulfils the criteria for a network intrusion dataset as it contains desirable features. With 225,745 packages and more than 80 features, the CIC IDS 2017 captured network activity for more than seven days over the course of five days of data collection (i.e., normal and intrusion). Brute force attack, heartbleed attack, botnet attack, DoS attack, web attack, and infiltration assault are among the seven categories that the attack simulation in the CIC 2017 dataset is broken down into.

3.3.2 CIC IDS 2018 (CSE-CIC-IDS2018) (01-03-2018) & (02-03-2018)

The CIC IDS 2018 dataset is an improvement of the CIC IDS 2017 dataset. More attack profiles were added. 80 network traffic features were captured in this dataset. The dataset is divided into various chunks by their date. This particular one used for this project was recorded on the 1st and 2nd of March 2018.

3.3.3 Feature Selection

As earlier stated, the CIC IDS2017 and CSE-CIC-IDS2018 datasets fulfill the criteria for a network intrusion dataset as they contain desirable attributes (Hamid et al., 2013). These attributes, excluding the protocol, are being considered because an increase in their values mostly points to a Distributed Denial of Service (DDoS) attack due to the flooding of the user's system. Some features were also converted into numerical format because of their relevance in detecting malicious traffic.

3.4 Performance Metrics

A "confusion matrix" was used to evaluate the performance of both the ensemble method and traditional classifiers. Accuracy, recall, precision, and the F1 measure were used as the major metrics. The minor metrics used are Specificity, MCC, KAPPA, AUC, FDR, FNR, FPR and NVP.

4 Results Obtained and Discussion

4.1 Data Analysis for Dataset 1

The "CIC IDS 2017 dataset" contains data for TOR and nonTor network traffic. Below is a countplot that shows a graphical view of the categories of data contained in the dataset. From the results obtained, it is clear that decision trees have the best accuracy score of 0.9746 while 0.9596 and 0.9579 are values obtained for accuracy for both Neural network and K Nearest Neighbour. An accuracy of 0.9455 was obtained for Support Vector Machine 0.9237 and 0.7066 were obtained for both Logistic Regression and Gaussian Naïve Bayes. The lowest value of accuracy of 0.1212 was obtained for the Quadratic Discriminant Analysis.

The results for the ensemble classifiers are as follows: the lowest accuracy of 0.9475 was obtained for Adaboost. Accuracy values of 0.9568 and 0.9675 were obtained for both Max Voting and Gradient Boosting. The best accuracy was obtained for the Random Forest ensemble with the accuracy value of 0.9828. From the foregoing, it is very clear that the Random Forest ensemble is most suitable for detecting any form of attack on the global network. On the other hand, Random Forest has the highest values for MCC and KAPPA of all four ensemble learning methods at 0.9191.

From Fig. 2, the confusion matrix shows that 11756 were correctly labeled as non-Tor network traffic and 1466 were correctly labeled as TOR traffic. 165 and 179 are known as Type I and Type II errors. This is because they were wrongly labeled. This model has a high accuracy of 97.56 and 89.88% precision. From Fig. 3, the confusion matrix shows that 11810 were correctly labeled as non-Tor network traffic and 1523 were correctly labeled as TOR traffic. 122 and 111 are known as Type I and Type II errors. This is because they were wrongly labeled. This model has a high accuracy of 97.56 and 89.88% precision.

Table 1. List of Datasets

Dataset	URL	Source	Remark
Dataset 1	https://www.unb.ca/cic/datasets/dos-dataset.html	University of New Brunswick, Canada	UNB IDS 2017
Dataset 2	https://registry.opendata.aws/cse-cic-ids2018/.		CSE-CIC-IDS2018
Dataset 3	https://registry.opendata.aws/cse-cic-ids2018/.		CSE-CIC-IDS2018

Table 2. Results of (CIC IDS2017) Dataset using Machine Learning Models and Ensemble Learning Models in Percentage form with emphasis on the Accuracy and Precision of the Models.

http://dx.doi.org/10.46792/fuoyejet.v8i4.1113 http://journal.engineering.fuoye.edu.ng/

Model	Accuracy	Precision	Recall	specificity	F1 Score	KAPPA	MCC	AUC	FDR	FNR	FPR	NVP
Decision Trees	97.46%	0.8988	0.8911	0.9861	0.8949	0.8805	0.8805	0.9386	0.1011	0.1088	0.0138	0.9850
K-Nearest Neighbor	95.79%	0.8500	0.7927	0.9807	0.8203	0.7965	0.7972	0.8867	0.1499	0.2072	0.0192	0.9716
Logistic Regression	92.37%	0.6703	0.7306	0.9504	0.6992	0.6556	0.6564	0.8405	0.3296	0.2693	0.0496	0.9623
Gaussian Naïve Bayes	70.66%	0.2910	0.9884	0.6677	0.4496	0.3227	0.4352	0.8280	0.7089	0.0115	0.3322	0.9976
Neural Network	95.96%	0.8909	0.7598	0.9871	0.8202	0.7976	0.8008	0.8735	0.1090	0.2401	0.0128	0.9675
Quadratic Discriminant Analysis	12.12%	0.1212	0.7580	1.0000	0.2162	0.0000	0.0000	0.5000	0.8787	0.0000	1.0000	nan
SVM	94.55%	0.8292	0.6936	0.9802	0.7553	0.7250	0.7286	0.8369	0.1707	0.3063	0.0197	0.9586
AdaBoost	94.75%	0.8176	0.7306	0.9775	0.7717	0.7422	0.7437	0.8541	0.1823	0.2693	0.0224	0.0224
Gradient Boosting	96.75%	0.8966	0.8279	0.9868	0.8609	0.8426	0.8434	0.9073	0.1033	0.1720	0.0131	0.9765
Random Forest	98.28%	0.9320	0.9258	0.9906	0.9289	0.9191	0.9191	0.9582	0.0679	0.0741	0.0093	0.9897
Max Voting	95.68%	0.8157	0.8316	0.9740	0.8236	0.7989	0.7990	0.9028	0.1842	0.1683	0.0259	0.9767

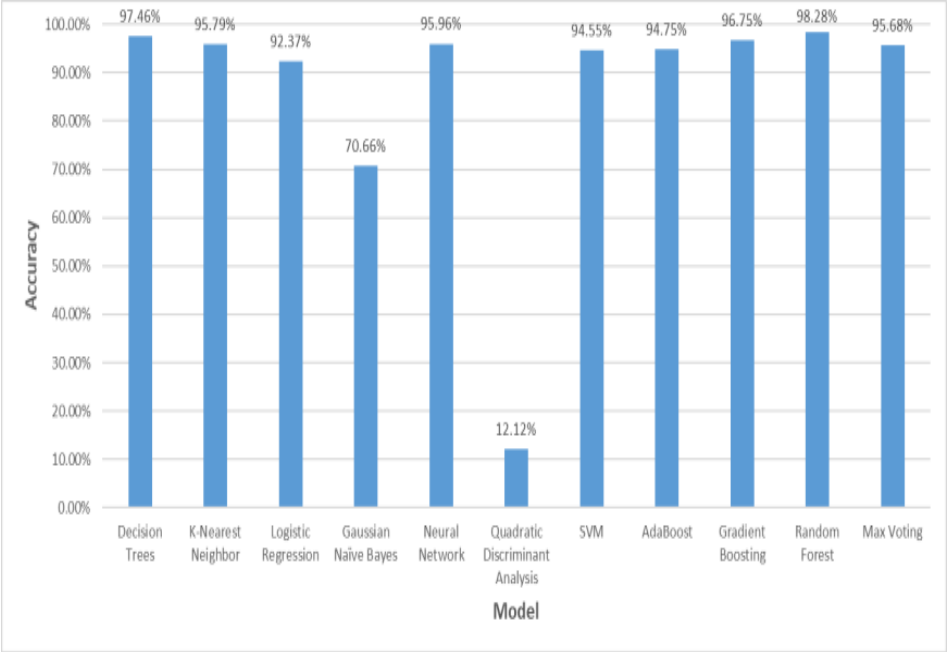

Fig. 1: Evaluation Metrics Versus Machine Learning Models and Ensemble learning models for the CIC IDS 2017 Dataset

Fig. 2: Confusion Matrix for Decision Trees

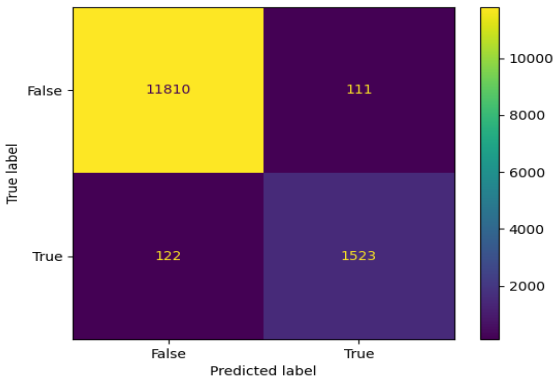

Fig 3: Confusion Matrix for Random Forest

Table 3. Results of (CSE-CIC-IDS2018) (02-03-2018) Dataset using Machine Learning Models and Ensemble Learning Models in Percentage form with emphasis on the Accuracy of the Models.

http://dx.doi.org/10.46792/fuoyejet.v8i4.1113 http://journal.engineering.fuoye.edu.ng/

Model	Accuracy	Precision	Recall	specificity	F1 Score	KAPPA	MCC	AUC	FDR	FNR	FPR	NVP
Decision Trees	99.86%	0.9966	0.9983	0.9987	0.9975	0.9965	0.9965	0.9985	0.0033	0.0016	0.0012	0.9993
K-Nearest Neighbor	99.81%	0.9951	0.9980	0.9981	0.9966	0.9953	0.9953	0.9981	0.0048	0.0019	0.0018	0.9992
Logistic Regression	93.93%	0.821	0.9986	0.9167	0.9011	0.8580	0.8665	0.9576	0.1789	0.0013	0.0832	0.9994
Gaussian Naïve Bayes	70.10%	0.4806	0.9989	0.5871	0.6489	0.4397	0.5303	0.793	0.5193	0.001	0.4128	0.9993
Neural Network	98.78%	0.9701	0.9866	0.9883	0.9782	0.9698	0.9699	0.9874	0.0298	0.0133	0.0116	0.9948
Quadratic Discriminant Analysis	94.63%	0.8384	0.9984	0.9264	0.9115	0.8734	0.8803	0.9624	0.1615	0.0015	0.0735	0.9993
SVM	93.48%	0.8098	0.9992	0.9102	0.8946	0.8482	0.8580	0.9547	0.1901	0.0007	0.0897	0.9996
AdaBoost	99.63%	0.9921	0.9948	0.9969	0.9934	0.9909	0.9909	0.9959	0.0078	0.0051	0.0030	0.9980
Gradient Boosting	99.82%	0.9953	0.9983	0.9982	0.9968	0.9956	0.9956	0.9982	0.0046	0.0016	0.0017	0.9993
Random Forest	99.89%	0.9970	0.9990	0.9988	0.9980	0.9972	0.9972	0.9989	0.0029	0.0009	0.0011	0.9996
Max Voting	94.03%	0.8230	0.9992	0.9178	0.9026	0.8601	0.8686	0.9585	0.1769	0.1769	0.0821	0.9996

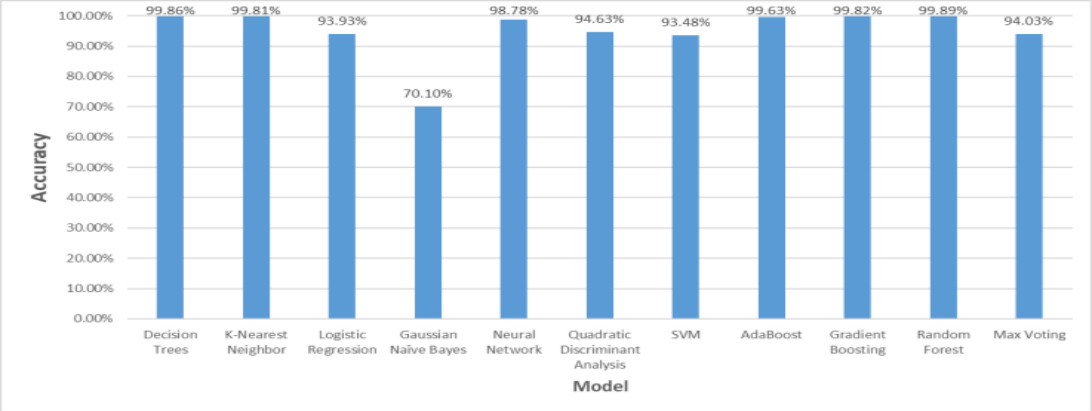

Fig. 4: Evaluation Metrics Versus Machine Learning Models and ensemble learning models for *(CSE-CIC-IDS2018)* Dataset

4.2 Analysis of Results for Dataset 2

The "CSE-CIC-IDS 2018" dataset of the 2nd of March2018 contains data on benign and malicious network traffic (Bot). Below is a countplot that shows a graphical view of the categories of data contained in the dataset. Amongst the regular machine learning algorithms used, the highest accuracy of 0.9986 was obtained for decision trees while the accuracy values of 0.9981 and 0.9878 were obtained for both K- Nearest Neighbor and Neural Network. An accuracy of 0.9463 was obtained for quadratic discriminant analysis while 0.9393 and 0.9348 were obtained as values of accuracy for logistic regression and SVM. From the experimentation, it is noted that Gaussian Nave Bayes had the lowest accuracy value of 0.7010.

The results of the Ensemble classifiers are as follows: Max Voting has 0.9403 level of accuracy, while Adaboost and Gradient Boosting, have the accuracy levels of 0.9963 and 0.9982, respectively. Random Forest ensemble has the best accuracy of 0.9989, making it the most appropriate for detecting malicious network traffic, although it worked fairly well in FDR, FNR, and FPR. On the other hand, Random Forest has the highest MCC and KAPPA of all the four ensemble learning methods at 0.9989.

From Fig. 6, the confusion matrix shows that 37646 were correctly labeled as benign network traffic and 14399 were correctly labeled as Bot traffic. 14 and 43 are known as Type I and Type II errors. This is because they were wrongly labeled. This model has a very high accuracy of 99.89% and a very high precision of 99.70%.

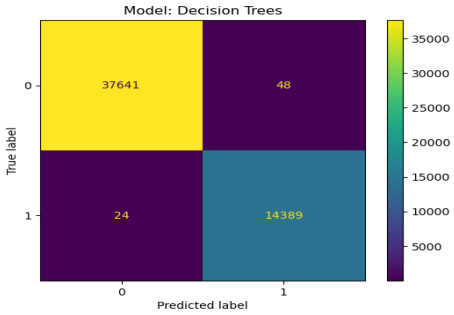

Fig. 5: Confusion Matrix for Decision Trees

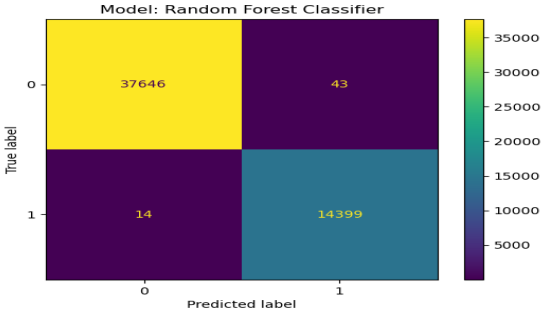

Fig. 6: Confusion Matrix for Random Forest

Table 4. Results of (CSE-CIC-IDS2018) (01-03-2018) Dataset using Machine Learning Models and Ensemble Learning Models in Percentage form with emphasis on the Accuracy of the Models.

http://dx.doi.org/10.46792/fuoyejet.v8i4.1113 http://journal.engineering.fuoye.edu.ng/

Model	Accuracy	Precision	Recall	specificity	F1 Score	KAPPA	MCC	AUC	FDR	FNR	FPR	NVP
Decision Trees	71.95%	0.3690	0.3020	0.8449	0.3321	0.1568	0.1581	0.5734	0.6309	0.6979	0.1550	0.8011
K-Nearest Neighbor	76.29%	0.4721	0.2243	0.9246	0.3041	0.1824	0.2008	0.5744	0.5278	0.7756	0.0753	0.7987
Logistic Regression	76.88%	0.2727	0.0005	0.9995	0.0010	0.0001	0.0020	0.5000	0.7272	0.9994	0.0004	0.7690
Gaussian Naïve Bayes	35.78%	0.2430	0.8418	0.2124	0.3771	0.0291	0.0572	0.5271	0.7569	0.1581	0.7875	0.8172
Neural Network	78.68%	0.7292	0.1226	0.9863	0.2099	0.1536	0.2376	0.5544	0.2707	0.8773	0.0136	0.7891
Quadratic Discriminant Analysis	27.49%	0.2365	0.9602	0.0690	0.3795	0.0141	0.0511	0.5146	0.7634	0.0397	0.9309	0.8525
SVM	77.08%	0.7194	0.0132	0.9984	0.0260	0.0178	0.0758	0.5058	0.2805	0.9867	0.0015	0.7711
AdaBoost	78.84%	0.8405	0.1039	0.9940	0.1849	0.1413	0.2479	0.5490	0.1594	0.8960	0.0059	0.7869
Gradient Boosting	79.16%	0.8125	0.1274	0.9911	0.2203	0.1682	0.2675	0.5593	0.1874	0.8725	0.0088	0.7908
Random Forest	77.69%	0.5446	0.2106	0.9471	0.3037	0.2007	0.2330	0.5788	0.4553	0.7893	0.0528	0.7997
Max Voting	77.98%	0.8495	0.0570	0.9969	0.1069	0.0801	0.1842	0.5270	0.1504	0.9429	0.0030	0.7787

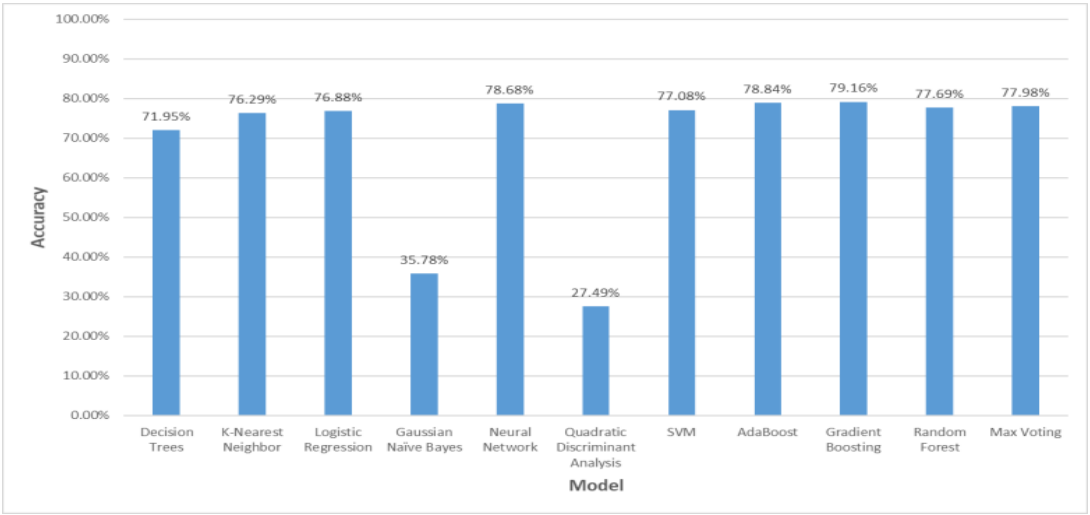

Fig. 7: Evaluation Metrics Versus Machine Learning Models and ensemble learning models for the *(CSE-CIC-IDS2018) (01-03-2018)* Dataset

4.3 Data Analysis for Dataset 3

The "CSE-CIC-IDS 2018" dataset of the 1st of March 2018 contains data on benign and infiltration network traffic. Below is a countplot that shows a graphical view of the categories of data contained in the dataset. Amongst the regular machine learning algorithms, neural networks have the best accuracy value of 0.7868 while support vector machines and logistic regression, accuracy values of 0.7708 and 0.7688. K-Nearest Neighbor has an accuracy of 0.7629, while decision trees and Gaussian Nave Bayes had an accuracy level of 0.7195 and 0.3578. It is observed that Quadratic discriminant analysis has the least accuracy value of 0.2749.

The results of the ensemble classifiers are: Random Forest has the lowest accuracy level at 0.7769 while Max Voting and Adaboost have accuracy values of 0.7798 and 0.7884. Gradient Boosting ensemble has the best value of accuracy of 0.9989. This feature has made it appropriate for detecting malicious network traffic, although it worked fairly well in FDR but poorly in FNR and FPR. On the other hand, gradient boosting has the highest MCC, while random forest has the highest KAPPA of all the four ensemble learning methods at 0.2007.

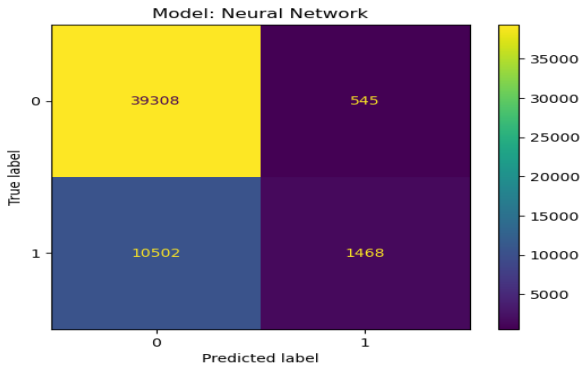

Fig. 8: confusion Matrix for Neural Network

From Fig. 8, the confusion matrix shows that 39308 were correctly labelled as benign network traffic and 1468 were correctly labelled as infiltration traffic. 10502 and 545 are known as Type I and Type II errors. This is because they were wrongly labelled. This model has an accuracy of 78.68% and a precision of 72.92%. This model has a "low false positive rate" and a very "high false negative rate.".

5 Conclusion

http://dx.doi.org/10.46792/fuoyejet.v8i4.1113 http://journal.engineering.fuoye.edu.ng/

In this paper, Intrusion Detection was discussed at length. Several papers were also reviewed to better put things in perspective. It is clear that Artificial Intelligence is able to mitigate intrusion detection to a reasonable extent if applied properly with the right tools. Also, this work has shown that firstly, Decision trees is one of the most effective Machine Learning Technique for Intrusion detection while the Random Forest ensemble learning model approach to malware classification, detection, and prevention is the most effective.

Furthermore, this paper also helps promote the use of Ensemble Learning models to improve the accuracy, precision and other metrics used in determining the usefulness and applicability of a machine learning model. Malware still poses a serious threat to users nonetheless, to curb the huge amount spent on remediating the effects of cyber-attacks on modern-day computer systems, adequate work has to be put into the detection and prevention of malware. Machine learning provides us with an effective approach to handling known and unknown types of malware attacks. It also provides a way of properly classifying malicious and benign network traffic data.

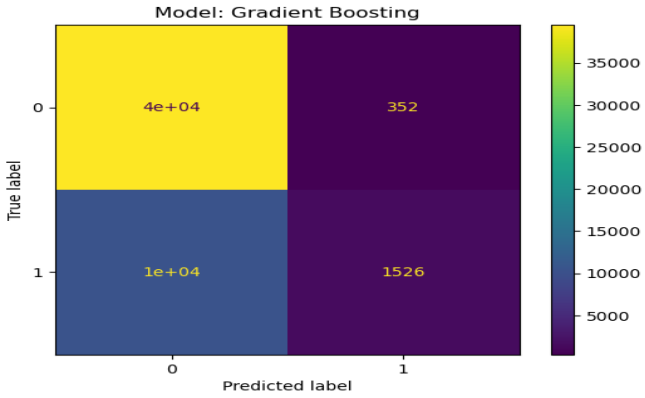

Fig. 9: Confusion Matrix for Gradient Boosting

From Fig. 9, the confusion matrix shows that 39308 were correctly labelled as benign network traffic and 1468 were correctly labelled as infiltration traffic. 10502 and 545 are known as Type I and Type II errors. This is because they were wrongly labelled. This model has an accuracy of 78.68% and a precision of 72.92%. This model has a low "false positive rate" and a very high "false negative rate.".

Despite the fact that the results of these tests were quite accurate, the level of efficacy is due to the environment where the tests were carried out. With the constantly changing threat landscape, the effectiveness of this approach may very well be obsolete or less applicable in the real world. This would be because of the introduction of new threats that would not be easily detected due to the lack of past data. A way to tackle this is to ensure that constant work is done in this field so that zero-day attacks can be easily detected and subsequently mitigated. To push the work further, an ensemble model of Deep learning is being proposed, apart from the hybridization of the algorithms used in this work, which is already in the pipeline.

REFERENCES

Abirami, A. and Palanikumar,S (2023) "BBBC-DDRL: A hybrid big-bang big-crunch optimization and deliberated deep reinforced learning mechanisms for cyber-attack detection". Computers and Electrical Engineering, Volume 109, Part B, 2023, 108773, ISSN 0045-7906. pp. 1-16.

Arif Yulianto, P. S. (2019). Improving AdaBoost-based Intrusion Detection System (IDS) Performance on CIC IDS 2017 Dataset. *The 2nd International Conference on Data and Information Science*.

Choi, J. K. (2008). A study on the detection of network reconnaissance attacks. *Proceedings of the 5th International Conference on Security and Cryptography*.

Bagaa, M., Taleb, T., Bernabe, J., & Skarmeta, A. (2020). A Machine Learning Security Framework for Iot Systems. *IEEE Access,* 114066–114077.

Chandrasekar, K., Cleary, G. , Cox, O. , Lau, H. , Nahorney, B. , Gorman, B.O. , O'Brien, D. , Wallace, S. , Wood, P. , Wueest, C. (2017). Internet Security Threat Report. *Technical Report, Symatec Corporation*.

Chao, S. W. (2015). CANN: an intrusion detection system based on combining cluster centers and nearest neighbors. *Knowl-Based Syst, vol. 78*, 13-21.

Chatterjee, S. and Hanawal, M.K. (2021). Federated Learning for Intrusion Detection in IoT Security: A Hybrid Ensemble Approach. *arXiv:2106.15349*.

Chen, Q. L. (2020). A deep learning based intrusion detection system for R2L attacks in wireless networks. *EEE Access, 8*, 146812-146820.

Cisco. (2023, February 16). *Intrusion Prevention Systems: Best Practices.* Retrieved from https://www.cisco.com/c/en/us/solutions/enterprise-networks/intrusion-prevention-best-practices.html

Cleary, G. P. (2018). Internet Security Threat Report. *Technical Report, Symantec Corporation*.

Dagon, D. (2006). Network intrusion detection. *Handbook of Information Security*.

F5 Networks. (2020). *F5 Essential App Protect*. Retrieved from F5 Essential App Protect: https://www.f5.com/products/f5-essential-app-protect

Gangwar, A. (2014). A survey on anomaly and signature based intrusion detection system. *Journal of Engineering Research and Applications*, 67-72.

Gibert, D. M. (2018). Using convolutional neural networks for classification of malware represented as images. *J. Comput. Virol. Hacking Tech. doi: 10.1007/s11416- 018- 0323- 0*.

Giriraj Vyas, S. M. (2014). Intrusion Detection Systems: A Modern Investigation. *International Journal of Engineering, Management & Sciences (IJEMS)*.

Golchha R., Joshi A. Gupta, G.P (2023) " Voting-based Ensemble Learning approach for Cyber Attacks Detection in Industrial Internet of Things". Procedia Computer Science, Volume 218, Pages 1752-1759, ISSN 1877-0509.

Goli Sushma, G. S. (2022). INTRUSION DETECTION SYSTEM USING MACHINE LEARNING TECHNIQUES. *International Journal of Engineering Technology Research & Management*.

Hamid, I. R. (2013). Using feature selection and classification scheme for automating phishing email detection. *Studies in Informatics and Control 22*, 61-70.

Hany Mohamed, H. H. (May 2018). Intrusion Detection System Using Machine Learning Approaches. *Egyptian Computer Science Journal Vol. 42 No.3*.

Hubballi. N., and Suryanarayanan. V (Aug 2014). False alarm minimization techniques in signature-based intrusion detection systems: A survey. *Comput. Commun, Vol. 49*, 1A17.

Iqbal H. Sarker, A. S. (2020). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*.

http://dx.doi.org/10.46792/fuoyejet.v8i4.1113 http://journal.engineering.fuoye.edu.ng/

Jayesh Zala, A. P. (2020). Intrusion Detection System using Machine Learning. *International Journal of Scientific Research in Computer Science,* Engineering *and Information Technology*.

Khan, R. Z. (2018). Analysis of ResNet and GoogleNet mod- els for malware detection. *J. Comput. Virol. Hacking Tech. doi: 10.1007/ s11416- 018- 0324- z*.

Khawaja, F., and Ali, N. (2021). R2L Attack Prevention Mechanism based on Least Privilege. *Procedia Computer Science, 180*, 206-213.

Khraisat A, Gondal I, Vamplew P. (2018). An anomaly intrusion detection system using C5 decision tree classifier. *Trends and applications in knowledge discovery and data mining*, 149-155.

Liu, S. C. (2022). A deep learning-based framework for detecting and mitigating DoS attacks. *IEEE Transactions on Network and Service Management, 19(1)*, 562-576.

Mohammadpour, L., Hussain, M., Aryanfar, A., Maleki R., and Sattar. F (2015). Evaluating Performance of Intrusion Detection System using Support Vector Machines: Review. International Journal of Security and Its Applications, pp.225-234. Mukherjee, D. (2023). "Detection of data-driven blind cyber-attacks on smart grid: A deep learning approach". Sustainable Cities and Society, Volume 92, 2023, 104475, ISSN 2210-6707. pp. 1-18.

Azeez, N.A and Odejinmi, S.O. (2023) "A Cyberstalking-Free Global Network with Artificial Intelligence Approach. Int. J. Information and Computer Security, Vol. 21, Nos. 1/2, 2023.

Azeez, N.A; Victor, O.E. and Sanjay, M. (2022) "Extracted Rule-Based Technique for Anomaly Detection in A Global Network" Int. J. Electronic Security and Digital Forensics, Vol. 14, No. 6, 2022

Azeez, N.A; Oladele, S.S; and Ologe, O. (2022). "Identification of Pharming in Communication Networks using Ensemble Learning" Nigerian Journal of Technological Development (UNILORIN), VOL. 19, NO.2.pp. 172-180.

Azeez, N.A; Ihotu A.M, Sanjay, M. (2021) "Adopting Automated White-List Approach for detecting Phishing Attacks" Elsevier Journal of Computers & Security 108 (2021) 102328, pp. 1-18.

Azeez, N.A; Idiakose, S.O; Onyema, C.J and Vyver, C.V (2021) "Cyberbullying Detection in Social Networks: Artificial Intelligence Approach" Journal of Cyber Security and Mobility, Vol. 10 4, 1–30. doi: 10.13052/jcsm2245-1439.1046

Azeez, N.A.; Odufuwa, O.E.; Misra, S.; Oluranti, J.;Damaševičius, R.(2021) Windows PE Malware Detection Using Ensemble Learning. Informatics 2021, 8, 10. https://doi.org/10.3390/informatics8010010

Okta. (2016). Retrieved from Okta: https://www.okta.com/identity-101/intrusion-prevention-system/

P.Akshaya. (2016). Intrusion Detection System Using Machine Learning Approach. *International Journal Of Engineering And Computer Science*, 18249-18254.

Paliwal, S. and Gupta, R. (2012). Denial-of-Service, Probing & Remote to User (R2L) Attack Detection using Genetic Algorithm. *Int. J. Comput. Appl*, 57–62.

Ployphan Sornsuwit & Saichon Jaiyen. (2019). A New Hybrid Machine Learning for Cybersecurity Threat Detection Based on Adaptive Boosting. *Applied Artificial Intelligence, 33:5*, 462-482.

Rendón-Segador, F.J., Álvarez-García, J.A., Angel Jesús Varela-Vaca, A.J. (2023). "Paying attention to cyber-attacks: A multi-layer perceptron with self-attention mechanism". Computers & Security, Volume 132, 2023, 103318, ISSN 0167-4048. pp. 1-13.

Shone, N. Ngoc, T.N Phai, V.D and Shi. Q (Feb 2018). A deep learning approach to network intrusion detection. *IEEE Trans. Emerg. Topics Comput. Intell., vol. 2, no. 1*, pp 41-50.

Sahingoz, O. Can and O. K. (2015). A survey of intrusion detection systems in wireless. *2015 6th international conference on modeling, simulation and applied optimization (ICMSAO)*, 1-6.

SANS Institute. (2023, February 16). Retrieved from Intrusion Prevention Systems: https://www.sans.org/white-papers/1065/

Vinayakumar, R., Alazab, M;. Soman, K;. Poornachandran, P;. Al-Nemrat, K and Venkatraman. S. (2019). Deep learning approach for intelligent intrusion. *IEEE Access, vol. 7*, 41525 - 41550.

Wagner, E. C. (2002). Code-Red: a case study on the spread and victims of an Internet worm. *Proceedings of the 2nd ACM SIGCOMM Internet Measurement Workshop*.

Warraich, Z.S. Morsi, W.G. (2023), "Early detection of cyber–physical attacks on fast charging stations using machine learning considering vehicle-to-grid operation in microgrids". Sustainable Energy, Grids and Networks 34 (2023) 101027. pp. 1-13.

Zhang, Y., Li, P., & Wang, X. (2019). Intrusion Detection for IoT Based on Improved Genetic Algorithm and Deep Belief Network. *IEEE Access* , 31711–31722.

Zhang, Z. (2019). Boosting Algorithms Explained, Theory, Implementation, and Visualization. *Towards Data Science*.

http://dx.doi.org/10.46792/fuoyejet.v8i4.1113 http://journal.engineering.fuoye.edu.ng/