Indonesian Journal of Data and Science

Volume 4 Issue 3
ISSN 2715-9930
https://doi.org/10.56705/ijodas.v4i3.88

Research Article

A Learning Approach for The Identification of Network Intrusions Based on Ensemble XGBoost Classifier

Oyelakin A.M [1,*], Akanbi M.B [2], Ogundele T.S [3], Akanni A.O [4], Gbolagade M.D [5], Rilwan M.D [6], Jibrin M.A [7]

[1] Al-Hikmah University, Ilorin, Nigeria, amoyelakin@alhikmah.edu.ng
[2] Kwara State Polytechnic, Ilorin, Nigeria
[3] Al-Hikmah University, Ilorin, Nigeria
[4] Al-Hikmah University, Ilorin, Nigeria
[5] Al-Hikmah University, Ilorin, Nigeria
[6] Kaduna State University, Kaduna, Nigeria
[7] Al-Hikmah University, Ilorin, Nigeria

Correspondence should be addressed to Oyelakin A.M; amoyelakin@alhikamah.edu.ng

Received 05 November 2023; Accepted 28 November 2023; Published 31 December2023

© Authors 2023. CC BY-NC 4.0 (non-commercial use with attribution, indicate changes).
License: https://creativecommons.org/licenses/by-nc/4.0/ — Published by Indonesian Journal of Data and Science.

Abstract:

The limitation of signature-based Intrusion Detection systems has given rise for the popularity of Machine learning (ML) approaches r for building such intrusion detection systems (IDSs). ML is a sub-filed of Artificial Intelligence that enables algorithms to learn from data and its applications have been widely accepted and used in many domains. To achieve a promising ML-based model that can identify attacks and intrusions in networks and the cyber space, different stages of machine learning approach like pre-processing, attribute selection, model building, hyper parameter tuning can be very important. CICIDS2017 intrusion dataset was used for all the experimentations. This study focuses on building cyber threat detection model based on the ensemble feature selection and classification method. Innovative approaches were used for the analysis and pre-processing of the dataset. Thereafter, XGboost algorithm was used for selecting relevant features from the default input attributes in each of the captures. Thereafter, the reduced features were employed in the identification of cyber intrusions. The average accuracy achieved in the 8 captures of the dataset is 98% while precision is 0.98. Also, recall is 0.98, f1-score is 0.98 while AUC ROC score is 0.99. The study concluded that XGBoost-based model was able to achieve promising results based on the proper dataset encoding, feature importance-based feature selection and tuning of the algorithm for intrusion identification.

Keywords: Intrusions, Machine Learning, Intrusion Identification, Model Classification Performance.
Dataset link: -

1. Introduction

An intrusion detection system (IDS) is used for network attacks identification which then alerts the administrator in good time. These IDSs are built using a wide variety of approaches. Machine Learning (ML) approaches have become popular for building such intrusion detection systems (IDSs). In ML field, algorithms learn from data and its applications have been found promising in security, medical and many other domains. The algorithms that can be used for building IDSs can be classified as single, hybrid and ensemble types. The chosen algorithms being used in this paper is an ensemble type. Ensemble classifiers became popular because they are targeted at improving the classification performance of a single classifier [1] as they combine several weak learners.

While building ML-based intrusion identification models, there is a need for dataset that contains representative attacks or intrusions so that the learning models can generalize well. In recent times, one of the most popular datasets that can be used for benchmarking attack classification models in CICIDS2017 dataset.As pointed out by [2] and [3] the CICIDS2017 dataset is better than some of the well-known IDS dataset such as KDD CUP99, NSL-KDD, Kyoto

2006, ISCX 2012 dataset as it contains benign and the most up-to-date common attacks. This study made use of the captures in the netflow file of the CICIDS2017 dataset. Also, [4] and [5] have argued that the CICIDS2017 dataset contains relatively new attacks types and are large representative for IDS studies when compared with old datasets

The limitation of signature-based Intrusion Detection systems has given rise for the popularity of Machine learning (ML) approaches r for building such intrusion detection systems (IDSs). ML algorithms have the ability to learn from data and its applications have been widely accepted in many domains. To achieve a promising ML-based model that can identify attacks and intrusions in networks and the cyber space, different stages of machine learning approach like data pre-processing, feature selection, model selection, hyper parameter tuning can be very important. This study focuses on building cyber intrusion detection model based on the ensemble feature selection and classification method with the use of XGBoost ensemble learner. Innovative approaches were used for the analysis and pre-processing of the dataset. Thereafter, XGboost feature importance was used for selecting relevant attributes from the available inputs in each of the captures. The reduced features were used for building in cyber intrusion model.

XGBoost provides parallel tree boosting in solving ML-based problems in a fast and accurate way. The exploratory analyses procedure includes: dataset description, computing the Statistical Summary, identification of the properties in the datasets. All the experimentations were carried out in Python programming language environment. Generally, feature selection aims at reducing the model complexity, reducing training time and building interpretable model [6], [7]. Thus, this work focuses on using feature importance approach for selecting promising features in the CICIDS2017 dataset and then use XGBoost ensemble learner for the classification of intrusions in the chosen dataset.

Related work

Built IDS from the NSL-KDD dataset by applying selected ML algorithms [8]. The algorithms belong to different categories and were trained and tested. The results of the ML techniques are discussed in details and authors argued that the results outperformed previous works. The dataset used for the experiment is considered to be older when compared with CICIDS2017 dataset. [9] proposed different scaling techniques for improving the performances of classification learners. Authors argued that there is performance difference between the classification algorithms based on the scaling technique used in most cases.

Conducted a comparative study between selected batch learning and data streaming classifier [10]. The authors' experimental results showed that data streaming algorithms achieved considerably higher performance in binary classification problems when compared with batch learning algorithms based on accuracy used as metrics. [11] carried out a systematic mapping and cross-benchmark evaluation of ensemble-based intrusion detection studies using different taxonomies. Furthermore, this study built a novel intrusion detection model that is based on stack of ensemble. The proposed approach uses parallel architecture to combine three individual ensemble learners in a homogeneous manner. The authors examined the performances of selected classification algorithms.

Carried out an analysis of some benchmark datasets that are used for building Network Intrusion Detection Systems [12]. The study generally described old and new datasets for IDS studies. However, it was observed that analyses were general and a detail report on a recent dataset named CICIDS2017 was not made. [6] used a filter-based approach named Information gain technique for attribute sub-set selection. The authors ranked and grouped the features using the minimum weight values and then built IDS models using Random five different machine classifiers. The study argued that RF-based classifier achieved the best performance accuracy of 99.86% out of all the learners.

Carried out an analysis of the CICIDS2017 dataset [2]. The study discussed some of the key features and components of the dataset. However, the study did not reveal some issues from the analysis and did not extend to the reporting of the open problems that are found in the dataset. [13] presented an AI technique for cyber-threats detection that is based on artificial neural networks. The authors used to use two benchmark datasets (NSL-KDD and CICIDS2017) and two datasets collected in the real world.

Similarly, carried out an analysis of CICIDS2017 dataset popularly used for building intrusion detection systems. The paper explored general characteristics of the dataset and mentioned some of the issues inherent to it without focusing on exploratory analyses and threat identification in it. [14] presented a hybrid approach utilizing the NSL-KDD dataset. The training dataset was 80% whereas the remaining 20% was used for testing on both binary and multi-class problems. Vote algorithm and information gain were applied for attribute selection. The detection accuracies of the two models were 99.81% and 98.56% respectively. However, the dataset used for the experimentation is old. Similarly, proposed build cyber intrusion detection model by making use of the fusion of chi square feature selection technique and multi class Support Vector Machine approach. [15] reviewed related studies on intrusion detection systems for a period between 2000 and 2007. The paper focused on studies that emphasised developing single, hybrid,

and ensemble techniques for intrusion identification. The study equally mentioned the achievements and drawbacks in ML-based IDS. Furthermore, [16] ML-based IDS in In-vehicle Network environment by using Remote Frame, The authors argued that the approach is novel since it used in-vehicle network environment for the experimentations.

Carried out an analysis of some intrusion detection datasets such as KDD-99, NSL-KDD and so on. The datasets used in the research were the ones that have been reported to have been very old. Conducted a statistical analysis on KDD CUP 99 dataset two main problems which highly affect the performance of intrusion detection systems built with it. Therefore, the authors proposed a data set named NSL-KDD that was meant to be an improvement of KDD CUP99 dataset and then attend to the mentioned shortcomings of the old dataset

2. Method:

Dataset

This work used an intrusion dataset named CICIDS2017 that was released by [2] and [3]. The dataset was collected from the Canadian Institute on Cyber Security dataset repository. The dataset was chosen because it is very large and contains several attacks and intrusion traces that are good for security studies.

Dataset Description

Pointed out that there are twenty-five (25) abstract behaviour of users that were built in the CICIIDS2017 dataset [3]. Thus, the dataset is based on the following protocols: HTTP, HTTPS, FTP, SSH, and email. The authors pointed out that the data was captured for varying period of time for five days. CICIDS2017 dataset contains both benign and non-benign behavioural patterns. Specifically, the dataset has a number of attacks such as: Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet, and DDoS [3]. During the dataset building, the demonstration of the attacks was carried out in the morning and afternoon on Tuesday, Wednesday, Thursday and Friday.

Methodological Process

Figure 1 is used to pictorially represent the various processes through which the cyber intrusions in the dataset can be efficiently classified.

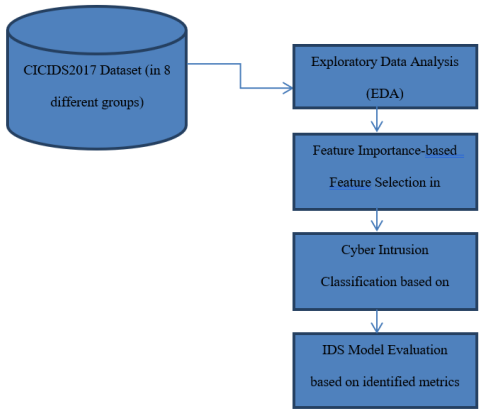


Figure 1. Machine Learning-based Methodological Process in the Study

The dataset is multi class in nature. For this reason, One Versus Rest (OVR) technique was used to tune it to a binary class type. This is to enable the chosen ensemble learning algorithm to classify the features in the dataset as being benign and malicious. All the experiments were conducted in Anaconda Python 4 environment using a 64-bit Intel(R) Core(TM) i7-7500U CPU with 16GB RAM (Windows 10 platform). After the EDA and feature encoding, each of the eight pre-processed captures in the dataset was saved in different folder and they were used for building the intrusion identification models.

Data Preprocessing

The numerical input features are of real and integer types while the target class is categorical type. Based on the flow patterns and data types in the dataset, the categorical features in the dataset were encoded. The essence of the feature encoding is to convert the categorical variables to numerical variables so as to make the features available to the machine learning algorithm in best usable format. Another step in the data cleaning process is the feature scaling so that the learning classifier will be able to learn adequately from scaled data. As part of the data pre-processing steps, min-max scaling technique was used to transform the features in the dataset. This data transformation approach is as argued. The Min-max Scalar is a data transformation technique that is used to transform an attribute scale and shifts its values along the X axis so that the transformed attribute ranges within the [0, 1] interval [8].

Feature Selection

In any ML-based classification problem, it is important to know which feature has more predictive power. Feature selection method is used to remove redundant features, helps in understanding the dataset, and reduces computing time, as well as improving model predictive performance [7]. Given the features in the CICIDS2017 dataset, XGBOOST feature importance is employed to achieve feature subset selection. Thirty-four attributes were selected based on their feature importance scores.

Model Building

Proposed XGBoost algorithm which is a supervised learning approach that is based on function approximation [17]. The algorithm is based on based on function approximation and it focuses on optimizing specific loss functions as well as applying several regularization techniques exploring different base learners. It involves the calculation of the value of the loss function for all those base learners. It uses an ensemble of decision trees and gradient boosting to make predictions. In this study, the algorithm is used for building the cyber intrusion identification model. The algorithm is built from gradient boosting decision tree algorithms as base learners. This algorithm is used in this case to classify intrusion in the dataset by passing though the different steps in algorithm 1. The dataset used for building the model is split in the ratio 85% and 15% as training and test ratios respectively. The selected metrics for the evaluation are Predictive Accuracy, Precision and Recall. The train test split ratio was used for the validation of how efficient the model is.

Algorithm 1:

//This algorithm contains procedures in the XGBooost classification of intrusion in CICIDS2017 dataset

Model Input: Dataset D: $D = \{(x_1, y_1), \ldots, (x_N, y_N), y_i \; belongs \; to \; R\}$

Model Output: Final intrusion classification: $f_m(x)$

//carry out the basic steps:

Initialization:

for m in 1,2,3,..., M

compute the gradient

compute the second derivative:

fit a new decision tree by reducing the loss function with regulation term:

find the best tree structure

find the minimal final loss

update the function $f_m(x)$:

Thereafter, produce the final intrusion identification model $f_m(x)$

3. Results and Discussion

Results of Exploratory Data Analysis

This study is making use of all the captures in the dataset as against the use of one of few captures. Based on the result of EDA, each capture in the dataset contains 78 input attributes and one class label. For instance, some of the attributed include: Destination Port, Flow Duration, Total Fwd Packets, Total Backward Packets, Total Length of Fwd Packets, Total Length of Bwd Packets and so on. It was also observed that each of the captures in the dataset has mixed feature types and this require innovative approach in handling prior to building intrusion classification model from the dataset. Also, it was observed that there are eight different captures in the dataset. Each of these captures contains attacks as recorded during the dataset building. Based on the period of capturing in those periods, the different captures in the dataset were renamed in this study as FriAfternoonPortScan, FriAfternoonDDOS, FriMorning, MonMorningHour, ThurAfternoonInfiltration, ThursdayMornWebAttacks, TueWorking, WedHour for easy referencing purposes. The details are as shown in **Table 1**.

Table 1. Description of each of the Captures in the Dataset

Capture Label Used	Capture Description in this Study
Capture 1	FriAfternoonPortScan
Capture 2	FriAfternoonDDOS
Capture 3	FriMorning
Capture 4	MonMorningHour
Capture 5	ThurAfternoonInfiltration
Capture 6	ThursdayMornWebAttacks
Capture 7	TueWorking
Capture 8	WedHour

Table 1 is used to represent the names assigned to the various captures of the CICIDS2017 Dataset used in the study. The experimental results obtained in **Table 2** are the true description of the attributes in the selected dataset.

Table 2. Dataset details based on EDA

Capture	Name Chosen for the Dataset capture	No of Input features	No of original samples/instances	No of unknown or missing values (data)	No of New Samples after Deletion
Capture 1	FriAfternoonPortScan	78	286,467	015	286452
Capture 2	FriAfternoonDDOS	78	225,745	004	225741
Capture 3	FriMorning	78	191,033	028	191005
Capture 4	MonMorningHour	78	529,918	064	539854
Capture 5	ThurAfternoonInfiltration	78	288,602	018	288584
Capture 6	ThursdayMornWebAttacks	78	170,366	020	170346
Capture 7	TueWorking	78	445,909	201	445708
Capture 8	WedHour	78	692,703	008	692695

Results of Model Evaluation for Intrusion Detection

Environment for the Experimentation

The environment where the model is built is a system with the following configuration: HP Core i7 system with 16GB main memory and 1TTerabyte Hard Disk Drive in a Windows 10 platform. Then, the identified metrics in the study are used for the experimental evaluation. For each of the models built from the eight captures of the CICIDS2017 dataset, the results of the metrics are as shown in **Table 3**.

Table 3. Experimental Results of Cyber Intrusion Classification Model based on XGBoost Ensemble

Dataset Capture	Learning Algorithm for Intrusion Classification	Accuracy Score (%)	Precision	Recall	F1-score	AUC-ROC	Comment on the Result
Capture 1	XGBoost Ensemble	99.00	0.97	0.99	0.98	0.99	The performances of the model across the five metrics are promising.
Capture 2	XGBoost Ensemble	98.00	0.98	0.98	0.98	1.00	The performances of the model across the five metrics are promising.
Capture 3	XGBoost Ensemble	98.00	0.98	0.97	0.98	0.99	The performances of the model across the five metrics are promising.
Capture 4	XGBoost Ensemble	99.00	0.98	0.98	0.98	0.99	The performances of the model across the five metrics are promising.
Capture 5	XGBoost Ensemble	98.00	0.98	0.98	0.98	0.99	The performances of the model across the five metrics are promising.
Capture 6	XGBoost Ensemble	98.00	0.97	0.98	0.98	0.99	The performances of the model across the five metrics are promising.
Capture 7	XGBoost Ensemble	98.00	0.97	0.98	0.98	0.99	The performances of the model across the five metrics are promising.
Capture 8	XGBoost Ensemble	97.00	0.98	0.98	0.97	0.98	The performances of the model across the five metrics are promising.

Benchmark Results Comparison

The results of the approach in this paper were compared with similar studies. On the average, it was observed that the ensemble learner achieved good performances for the classification of intrusions in the chosen CICIIDS2017 dataset compared to similar studies. For instance, the XGBoost-based model achieved promising classification results

in all the eight captures used for the experimentation. The models built from the eight different captures performed generally better than the one in similar studies.

Discussion of Results

The basic characteristics of the CICIDS2017 dataset were revealed from the exploratory analyses of the dataset captures. Thereafter, this study introduced the basic steps followed in building the ensemble model. Then, the work introduced how ML approaches are widely being used for building intrusion detection systems as against the signature-based intrusion detection approaches. The focus is on building intrusion detection model that is based on the feature importance for promising attribute selection and classification with the use of XGBoost ensemble learner. Five metrics namely: accuracy, precision, recall, f1-score and AUC score are used for the evaluation. The study achieved improved performances across the five selected metrics in the eight captures of the dataset. For instance, the average accuracy achieved in the 8 captures of the dataset is 98%, precision is 0.98, recall is 0.98, f1-score is 0.98 while AUC score is 0.99. Thus, it can be argued that the results performed by XGBoost-based model for the identification of intrusion in the chosen dataset are promising and better some of the similar studies. It is equally observed that our approach is more detailed as all the captures were used for the experimental analyses at each stage of the study.

The application of the Gradient Boosting Classifier on the MangoLeafBD dataset for the classification of Powdery Mildew and Sooty Mould yielded notable results. The performance metrics, evaluated through 5-fold cross-validation, demonstrated variability across different folds. The accuracy scores ranged from 0.63 to 0.73, with the highest being in the first fold. Precision metrics were consistently higher, ranging from 0.78735632 to 0.82467532, indicating a strong likelihood that the predicted positive cases were indeed positive. Recall scores paralleled the accuracy scores, which is expected as they both reflect the model's ability to correctly identify positive cases. The F1-scores, which balance precision and recall, varied between 0.57131271 and 0.70876928, reflecting some fluctuation in the model's overall performance. The detailed results are presented in **Table 1** for a clearer understanding and comparison of the metrics across different iterations

4. Conclusion

This study introduced CICIDS2017 dataset and how ML approaches are promising for building improved intrusion detection systems as against the signature-based Intrusion detection approaches. The study argued that to achieve a promising ML-based model that can identify attacks better in networks and the cyber space, different stages of machine learning approach like data pre-processing, feature selection, model building and hyper parameter tuning can be very important. This work used various approaches at the data cleaning, feature selection and classification stages. Then, XGBoost ensemble learner was used for the intrusion classification models in the eight captures of the dataset. The dataset was first analysed and pre-processed. Thereafter, XGboost algorithm was used for selecting relevant features and for identification of intrusions. The study achieved improved performances across the five selected metrics in the eight captures of the dataset. The efficiency is measured by the way the algorithm is able to learn from all the captures in the dataset. The average performances of the models were then computed.

References:

[1] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, “On combining classifiers,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998, doi: 10.1109/34.667881.

[2] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward generating a new intrusion detection dataset and intrusion traffic characterization,” in *ICISSP 2018 - Proceedings of the 4th International Conference on Information Systems Security and Privacy*, 2018, vol. 2018–January, pp. 108–116, doi: 10.5220/0006639801080116.

[3] I. Sharafaldin, A. Gharib, A. H. Lashkari, and A. A. Ghorbani, “Towards a Reliable Intrusion Detection Benchmark Dataset,” *Softw. Netw.*, vol. 2017, no. 1, pp. 177–200, 2017, doi: 10.13052/jsn2445-9739.2017.009.

[4] B. A. Tama, L. Nkenyereye, S. M. R. Islam, and K.-S. Kwak, “An Enhanced Anomaly Detection in Web Traffic Using a Stack of Classifier Ensemble,” *IEEE Access*, vol. 8, pp. 24120–24134, 2020, doi: 10.1109/ACCESS.2020.2969428.

[5] Y. Zhang, X. Chen, L. Jin, X. Wang, and D. Guo, “Network Intrusion Detection: Based on Deep Hierarchical Network and Original Flow Data,” *IEEE Access*, vol. 7, pp. 37004–37016, 2019, doi:

10.1109/ACCESS.2019.2905041.

[6] Kurniabudi, D. Stiawan, Darmawijoyo, M. Y. Bin Idris, A. M. Bamhdi, and R. Budiarto, "CICIDS-2017 Dataset Feature Analysis With Information Gain for Anomaly Detection," *IEEE Access*, vol. 8, pp. 132911–132921, 2020, doi: 10.1109/ACCESS.2020.3009843.

[7] Akinyemi Moruff Oyelakin and Jimoh Rasheed G, "A Survey of Feature Extraction and Feature Selection Techniques used in Machine Learning-Based Botnet Detection Schemes."

[8] Y. Almutairi, B. Alhazmi, and A. Munshi, "Network Intrusion Detection Using Machine Learning Techniques," *Adv. Sci. Technol. Res. J.*, vol. 16, no. 3, pp. 193–206, Jul. 2022, doi: 10.12913/22998624/149934.

[9] L. B. V. de Amorim, G. D. C. Cavalcanti, and R. M. O. Cruz, "The choice of scaling technique matters for classification performance," *Appl. Soft Comput.*, vol. 133, p. 109924, Jan. 2023, doi: 10.1016/j.asoc.2022.109924.

[10] K. S. Adewole *et al.*, "Empirical Analysis of Data Streaming and Batch Learning Models for Network Intrusion Detection," *Electronics*, vol. 11, no. 19, p. 3109, Sep. 2022, doi: 10.3390/electronics11193109.

[11] B. A. Tama and S. Lim, "Ensemble learning for intrusion detection systems: A systematic mapping study and cross-benchmark evaluation," *Comput. Sci. Rev.*, vol. 39, p. 100357, Feb. 2021, doi: 10.1016/j.cosrev.2020.100357.

[12] M. Ghurab, G. Gaphari, F. Alshami, R. Alshamy, and S. Othman, "A Detailed Analysis of Benchmark Datasets for Network Intrusion Detection System," *Asian J. Res. Comput. Sci.*, pp. 14–33, Apr. 2021, doi: 10.9734/ajrcos/2021/v7i430185.

[13] J. Lee, J. Kim, I. Kim, and K. Han, "Cyber Threat Detection Based on Artificial Neural Networks Using Event Profiles," *IEEE Access*, vol. 7, pp. 165607–165626, 2019, doi: 10.1109/ACCESS.2019.2953095.

[14] S. Aljawarneh, M. Aldwairi, and M. B. Yassein, "Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model," *J. Comput. Sci.*, vol. 25, pp. 152–160, Mar. 2018, doi: 10.1016/j.jocs.2017.03.006.

[15] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin, "Intrusion detection by machine learning: A review," *Expert Syst. Appl.*, vol. 36, no. 10, pp. 11994–12000, Dec. 2009, doi: 10.1016/j.eswa.2009.05.029.

[16] H. Lee, S. H. Jeong, and H. K. Kim, "OTIDS: A Novel Intrusion Detection System for In-vehicle Network by Using Remote Frame," in *2017 15th Annual Conference on Privacy, Security and Trust (PST)*, Aug. 2017, pp. 57–5709, doi: 10.1109/PST.2017.00017.

[17] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.