Indonesian Journal on Data Science
ISSN 2987-7423 - Vol. 2, No. 1, May 2024, hlm 1-9

Tree-based Machine Learning Ensembles and Feature Importance Approach for the Identification of Intrusions in UNR-IDD Dataset

Oyelakin A. M.[a,1,*], Saka B. A.[a,2,], Bakare-Busari Z.M.[a,3,], Olaleye O. J.[b,4,], Akee S. A.[a,5,], Yisa F. I.[a,6,], Lasisi I. O.[a,7,], Adam A. A.[a,8,], Adebiyi-Sodunke A. A.[a,9]

[a]Department of Computer Science, Crescent University, Abeokuta, Nigeria
[b]Department of Computer Science, Bells University of Technology, Ota, Nigeria
[1]moruff.oyelakin@cuab.edu.ng*; [2]badirat.saka@cuab.edu.ng; [3]bakare.busari@cuab.edu.ng; [4]ojolaleye@bellsuniversity.edu.mg; [5]sakirat.akee@cuab.edu.ng; [6]al.fattah2003@gmail.com; [7]ismail.lasisi@cuab.edu.ng ; [8]adenike.adam@cuab.edu.ng; [9]adebiyi1318@cuab.edu.ng

* corresponding author

ABSTRACT

Detection of intrusions from network data with the use of machine learning techniques has gained great attention in the past decades. One of the key problems in the network security domain is the availability of representative datasets for testing and evaluation purposes. Despite several efforts by researchers to release datasets that can be used for benchmarking attack detection models, some of the released datasets still suffer from one limitation or the other. Thus, some researchers at the University of Nevada released a new dataset named UNR-IDD dataset which was argued to be free from some of the limitations of the past datasets. This study used Tree-based ensemble approaches for building binary intrusion identification models from the UNR-IDD dataset. Decision Tree algorithms are used as base classifiers in the Extra Trees, Random Forest and AdaBoost-based intrusion detection models. The results of the experimental analyses carried out indicated that the three ensembles performed excellently when feature selection was used compared to when all features were applied. For instance, Extra Trees model achieved an accuracy of 0.97, precision of 0.98, recall of 0.98 and f1-score of 0.98. Similarly, Random Forest model achieved an accuracy of 0.98, precision of 0.98, recall of 0.99 and f1-score of 0.98. Adaboost-based model had an accuracy of 0.96, precision of 0.96, recall of 0.99 and f1-score of 0.98. It was deduced that Random Forest intrusion classification model achieved slight overall best results when compared to the other models built. It is concluded that the three homogeneous ensemble models achieved very promising results while feature importance was used as attribute selection method compared to when no feature selection technique is used.

ARTICLE INFO

Article history
Received: 6 Februari 2024
Revised: 20 March 2024
Accepted: 11 May 2024

Keywords
Intrusion Detection Datasets,
Intrusion Detection,
Tree-based Algorithms, Ensemble Learners

1. Introduction

Intrusion Detection System is a system that monitors network traffic for suspicious activity and issues alerts when such activity is discovered [1]. Globally, computer networks are constantly being attacked by people with bad intent. Several benchmark intrusion detection datasets have been widely used for building intrusion detection systems by applying a wide range of innovative techniques. Examples of some of these datasets are: KDDCUP 99, NSL-KDD, Kyoto 2006, ISCX2016, DARPA, ADFA-LD, UNSW-NB15, CICIDS2017, CICIIDS2018 and many others. However, authors in [2] argued some of some of these datasets have several drawbacks and thus the intrusion detection systems built from them do suffer from sub-optimal performance and inadequate tail class representations.

Several past studies have used machine learning approaches for the identification of intrusions in different past network datasets. In some of such intrusion detection studies, synthetic datasets were used. In some other ones, the datasets have smaller samples, while some of the intrusion datasets are filled with other limitations. Thus, the performances of the intrusion detection models may be badly affected if the shortcomings in the datasets are not properly and adequately handled.

Therefore, [2] built a new intrusion detection dataset that is targeted at getting rid of some of the limitations in some of the previous intrusion detection datasets. The word UNR-IDD was derived from University of Nevada - Reno Intrusion Detection Dataset (UNR-IDD) that provides researchers with a wider range of samples and scenarios [2]. The authors of UNR-IDD dataset argued that the dataset is an improvement having observed the deficiencies of some of the existing intrusion detection datasets. Authors further mentioned that the dataset provides researchers with a wider range of attack samples and scenarios and it contains the following kinds of attacks: TCP-SYN Flood attack, Port Scan, Flow Table Overflow, Blackhole and Traffic Diversion Attack. Since its release, the UNR-IDD intrusion detection dataset is gradually being used for benchmarking intrusion detection models. The dataset is based on the use of Network Port Statistics. To the best of the researchers' knowledge, there are very few works that have used the newly released UNR-IDD dataset for benchmarking machine learning-based intrusion detection models.

Machine learning algorithms have been found to be very promising when it comes to classification problems and in business applications [3]. Aside single learners, there are ensemble algorithms that are more promising for classification or regression problems ([4];[5]). On the basis of this, authors in [6] have demonstrated how ensemble learners can be very promising for the identification of phishing websites. Experimental results in the study showed how phishing websites were effectively classified using the chosen ensemble algorithms. This current study focuses on building intrusion detection models that are based on Tree ensemble learning algorithms using the newly released intrusion detection dataset. This study further seeks to investigate how Feature Importance as an attribute selection method can influence the performances of the selected Tree-based ensemble intrusion detection models built.

2. Related Work

Authors in [2] proposed models for the identification of attacks in the newly released UNR-IDD dataset. Authors built machine learning models for the classification of intrusions in the new benchmark dataset with a view to showing how efficient performances were arrived at in the classification tasks. Authors reported that their evaluation results showed that UNR-IDD is better than existing NIDS datasets with an Fµ score of 94% and a minimum F-score of 86%. Researchers in [7] built intrusion classification model that is based on XGBoost algorithm. The authors made use of the ensemble algorithm for the identification of intrusions in the chosen CICIDS2017 dataset. It was argued that the approach achieved better performances using accuracy, precision, recall, f1-score and AUC ROC.

Authors in [8] built machine learning based intrusion detection models with the use of NSL-KDD dataset for benchmarking. The authors made use of Support Vector Machine, J48, Random Forest, and Naïve Bayes algorithms for the model building. They argued that the results were promising. However, the dataset is very old and it contains some other limitations.[9] proposed the use of single and ensemble learners for the identification of phishing attacks. The study focused on the identification of phishing-related threats with the use of single and ensemble learning algorithms respectively. The experimental results showed that the ensemble approach achieved better performances in all the metrics compared to the single learner.

[10] proposed a Pigeon Inspired Optimizer based feature selection approach for building an Intrusion Detection System. The authors made use of three selected dataset for evaluation. The datasets include: KDDCUP99, NLS-KDD and UNSW-NB15. They further made the conclusion the proposed method outperformed several feature selection algorithms from state-of-the-art related works in

terms of TPR, FPR, accuracy, and F-score.[11] built an intrusion detection system using UNSW-NB15 dataset. The study emphasised the impact of feature selection in achieving improved IDS. The authors argued that their approach achieved promising results.

[12] came up with a Genetic Algorithm (GA) and Logistic Regression (LR) wrapper-based feature selection method for intrusion detection. The researchers made use of UNSW-N15 and KDDCUP99 datasets. They argued that the results showed that the GA-LR coupled with the Decision Tree algorithm attained a detection score of 81.42% as well as a FAR of 6.39% with 20 features out the 42 features present within the UNSW-NB15 feature space. With regards to the KDDCup99 dataset, the GA-LR in conjunction with the DT classifier obtained a detection score of 99.90% and a FAR rate of 0.105% using 18 features.

3. Method

3.1 Proposed Method

This study made use of ensemble learning approaches for building improved intrusion classion models. The work used a new intrusion dataset named UNN-IRD. In all our experiments, focus was on the binary classification of intrusions in the dataset. The key stages involved in the proposed method are as captured in figure 1. The study assumes a binary classification problem.

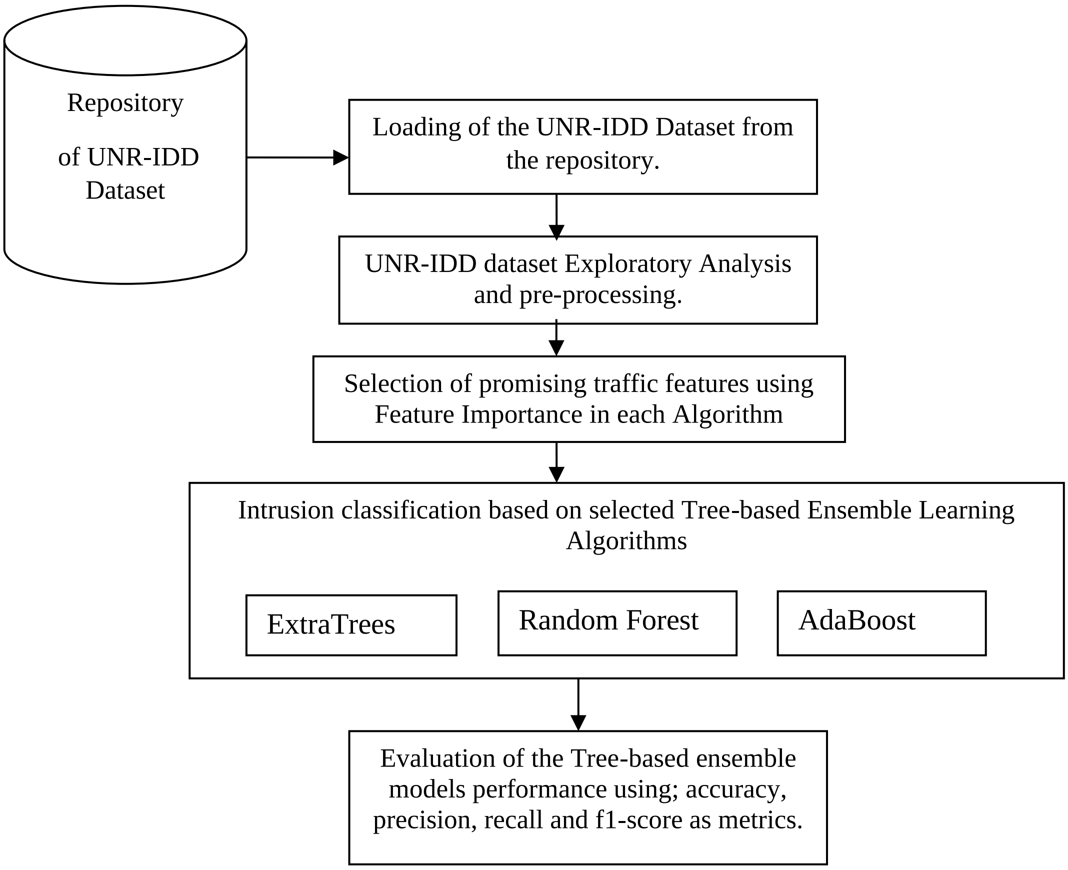
Repository
of UNR-IDD
Dataset
Loading of the UNR-IDD Dataset from the repository.
UNR-IDD dataset Exploratory Analysis and pre-processing.
Selection of promising traffic features using Feature Importance in each Algorithm
Intrusion classification based on selected Tree-based Ensemble Learning Algorithms
ExtraTrees
Random Forest
AdaBoost
Evaluation of the Tree-based ensemble models performance using; accuracy, precision, recall and f1-score as metrics.

Fig. 1. Methodological Flow of the Proposed Tree-based Ensemble Models for Attack Detection

Ensemble learning algorithms operate by combining the decisions from multiple single learners so as to improve the overall performance of a model [13]. The ensemble classifiers used in this study are Extremely Randomised Trees (Extra Trees), Random Forest (RF) and AdaBoost algorithms. The Extra Trees algorithm constructs multiple trees in similar manner to RF algorithm during training time. However, during the training, the former algorithm builds trees over every sample in the dataset but with different subsets of features. Aside this, RF algorithm is a type of ensemble algorithm that is based on bagging while AdaBoost is based on boosting. AdaBoost iteratively trains weak learners and calculates a weight for each one, and this weight represents the strength of the weak learner [14]. Adaptive Boosting was proposed by [15] and it was argued that this type of ensemble algorithm is very promising in classification problems. Table 1, 2 and 3 are used to describe the chosen ensemble machine learning algorithms used in detail. The training and testing split ratios used for the validation of the models is 80:20 respectively when the experiment is run for each of the algorithms.

Table 1. Algorithm 1: Extra Trees Algorithm for Intrusion Classification

Input: Set of input features in the UNR-IDD Dataset
Output: Binary Classification of Intrusions based on Extra Trees
1. Set the Base Learner=Decision Trees 2. Set the Number of Learning Rounds 3. Create many decision trees from the features in the dataset, 4. Perform sampling for each tree randomly, without replacement. 5. Perform random selection of a splitting value for each feature 6. Randomly selects a split value and classify intrusions in the dataset 7. Iterate until the best intrusion classification results are obtained

Table 2. Algorithm 2: Random Forest Algorithm for Intrusion Classification

Input: Set of input features in the UNR-IDD Dataset
Output: Binary Classification of Intrusions based on Random Forest
1. Set Base Learner=Decision Trees 2. Set the Number of Learning Rounds 3. Train each decision tree on a different subset of the intrusion dataset as per the hype parameter, 4. Classify Intrusions based on Decision Tree Splitting 5. Iterate until the best classification results are obtained 6. Produce the final classification of intrusions from all the trees based on averaging

Table 3. Algorithm 3: AdaBoost Algorithm for Intrusion Classification

Input: Set of input features in the UNR-IDD Dataset
Output: Binary Classification of Intrusions based on AdaBoost
1. Set Base Learner=Decision Trees 2. Set the Number of Learning Rounds 3. Assign equal weights to all the data points 4. Train a weak learner using the current sample weights 5. Calculate the error of the weak classifier 6. Calculate the weight of the weak classifier based on the error 7. Update the sample weights based on the weak classifier's performance 8. Normalize the sample weights 9. Combine the weak classifiers using a weighted majority vote. 10. Iterate until conditions that give better performance are met

For each of the intrusion detection models being built in this study, the ensemble learning algorithms were properly tuned until the best performances were recorded.

3.2 Dataset Collection

The dataset used in this study was collected from https://www.tapadhirdas.com/unr-idd-dataset.It is a dataset that was released by [2].

3.3 Dataset Description and Exploratory Analysis

The UNR-IDD dataset was built at the University of Nevada by [2]. The full name of the dataset is University of Nevada Reno Intrusion Detection Dataset (UNR-IDD). It uses network port statistics. The authors claimed that the intrusion types were selected for this dataset as they are common cyber-attacks that can occur in any networking environment. Authors in [2] have also argued that the UNR-IDD dataset is free from missing values. The authors equally pointed out that the dataset has both binary and multi-class labels which enable researchers to be a able to build binary-based intrusion and multi-class attack detection models respectively. The exploratory analysis carried out revealed that the dataset has few categorical data types that were encoded as part of the data pre-processing in this study. The EDA carried out also established that there are four ports in the network used for the traffic collection while building the dataset. Lastly, the EDA showed that there are various features and the data types in the dataset. We have features such as *Switch ID, Port Number, Received Packets Received Bytes, Sent Bytes, Sent Packets, Port alive Duration (S),Packets Rx Dropped, Packets Tx Dropped, Packets Rx Errors, Packets Tx Errors, Delta Received Packets, Delta Received Bytes, Delta Sent Bytes, Delta Sent Packets, Delta Port alive Duration (S), Delta Packets Rx Dropped, Delta Packets Tx Dropped, Delta Packets Rx Errors, Delta Packets Tx Errors, Connection Point, Total Load/Rate, Total Load/Latest, Unknown Load/Rate,Unknown Load/Latest, Latest bytes counter, is_valid, Table ID, Active Flow Entries, Packets Looked Up , Packets Matched, Max Size, Label and Binary Label.*The importance of Exploratory Data Analysis in ML-based security researches was further established by the authors in [16].

3.4 Dataset Pre-Processing

From the exploratory data analysis, it was observed that the dataset has many numerical and few categorical data. Thus, the categorical features have to be encoded so that the Tree-based algorithms can use them for the classification of intrusion evidence.

3.5 Feature Selection Approach

Feature selection is generally used to improve the performances of classification models as and it has been widely used in many domains [17]. In the three models being built, feature importance is used for selecting promising attributes in the UNR-IDD dataset. That is, we obtained feature importance scores from the tree-based models. The focus is to reduce the feature space and build intrusion detection models that will be less complex, have reduced training time and achieve high rate of attack detection.

3.6 Model Evaluation

The metrics used for the evaluation of the models include: Accuracy, Precision, Recall and F1-Score. For evaluating the performances of the three models based on how best they are able to classify intrusions the values obtained for the metrics in each of the learning algorithms were recorded.

4. Results and Discussion

4.1 Classification Algorithms and Model Building

For each of the network traffic in the dataset, the focus is to correctly classify the traffic as either intrusion or non-intrusion using the selected Tree-based ensemble learners. Tree-based methods have become one of the most flexible, intuitive, and powerful data analytic tools for exploring complex data structures [18]. The classification algorithms chosen in this study are Extra Trees, Random Forest and Adaboost. Thus, the problem is focused on investigating how the selected Tree-based homogeneous ensembles perform in the classification of intrusions in the UNR-IDD dataset, given the set of features and samples in it. The training test split ratio used in each of the experimentations is 80:20.

4.2 Experimental Results

4.2.1 UNR-IDD Dataset Exploratory Analysis

As argued in a paper by [16], this study carried out exploratory analysis of the UNR-IDD dataset so as to understand the data better before using it to build intrusion classification models.

Table 4. Result of Dataset Description based on the Exploratory Analysis

	SwitchID	Port Number		Label	Binary Label
0	of:000000000000000c	Port#2	…	5	0
1	of:000000000000000c	Port#3	…	5	0
2	of:000000000000000c	Port#4	…	5	0
3	of:000000000000000a	Port#1	…	5	0
	…	…	…	…	…
37406	of:0000000000000006	Port#2	…	4	0
37407	of:0000000000000006	Port#3	…	4	0
37408	of:0000000000000009	Port#1	…	4	0
37409	of:0000000000000009	Port#2	…	4	0
37410	of:0000000000000009	Port#3	…	4	0

Table 5. Summary Statistics of the Dataset based on the Exploratory Analysis

	Received Packets	Received Bytes	…	Label	Binary label
count	37411.000000	3.741100e+04	…	37411.000000	37411.000000
mean	21618.897169	2.647491e+07	…	2.663174	0.100853
Std	65283.170126	3.703044e+07	…	1.959171	0.301138
Min	9.000000	7.860000e+02	…	0.000000	0.000000
25%	329.000000	9.104050e+04	…	1.000000	0.000000
50%	1170.000000	1.263052e+07	…	3.000000	0.000000
75%	3417.000000	3.783230e+07	…	4.000000	0.000000
max	352772.000000	2.715925e+08	…	5.000000	1.000000

Table 4 shows the summary of the features and samples in the dataset. The table 5 provides the statistical summary of the features in the dataset. As part of the exploratory analysis in the study, it was observed that there are 37,411 samples and 33 input attributes and one class label. Also, the dataset contains no missing values, and it has three different kinds of data types. For instance, there are: bool, int64, and object data types.It was revealed that there are two duplicate records. The duplicates were handled as part of the per-processing steps in the study.

4.2.2 Feature Encoding and Scaling

The findings of the exploratory analysis in this study enabled the researchers to convert the following categorical features: SwitchID, Port Number,is_valid, in the dataset using one-hot encoding technique. Thereafer, the features in the dataset were normalized using scaling. The importance of these techniques is to make the features well utilised by the classification algorithms.

4.2.3 Results of Selected Features

Each of the tree-based classification ensemble algorithms were used for the selection of promising features based on feature scoring. The features were ranked in each of the algorithms based on their scores. Then, in each of the algorithms, different features were selected based on the threshold set. For the Random Forest Algorithm, twenty-six (26) features were selected. For the Extra Trees Algorithm, twenty-four (24) features were selected while seventeen (17) features were selected in the Ada Boost algorithm. All the selected features are based on the assumption that they have promising scores compared to others whose scores are very low. The features were used while building the intrusion classification models.

4.3 Results of Classification of Intrusions based with and Without Feature Selection

Authors in [19] have emphasised the promises of feature engineering in machine learning classification tasks. This study established how the chosen feature selection method influenced the performances of RF, Extra Trees and AdaBoost algorithms in the intrusion classification task.

Table 6. Results of the Ensemble-based Intrusion Classification without Feature Selection

Metric/Classification Algorithm	Extra Trees	Random Forest (RF)	AdaBoost
Accuracy	0.61	0.62	0.40
Precision	0.59	0.60	0.50
Recall	0.58	0.59	0.40
f1-Score	0.59	0.60	0.36

Table 7. Results of the Ensemble-based Intrusion Classification with Feature Selection

Metric/Classification Algorithm	Extra Trees	Random Forest (RF)	AdaBoost
Accuracy	0.97	0.98	0.96
Precision	0.98	0.98	0.96
Recall	0.98	0.99	0.99
f1-Score	0.98	0.98	0.98

4.4 Discussion of Results

The exploratory analysis carried out on the dataset revealed more information about the characteristics of the dataset. The analysis guided the approach being used to use the dataset for better classification of intrusions in it. For instance, it was revealed that there are both numerical and few categorical data in the dataset chosen for the study. Thus, the EDA provided the some of the good sides of the network data as well as the justification for encoding the categorical attributes so that the chosen learning algorithms will be able to process all the features better. That is, for the features to be in the right usable format for the learning algorithms, the categorical (textual) features were encoded. Label encoding was used to convert the few categorical features into numeric values. In each of the learning ensembles, Feature importance was used to select promising attributes that were used for training intrusion classification models. For the intrusion classification, experimental results showed that all the three algorithms performed excellently when feature importance was used for selecting promising features. The performances of the three models when feature selection was not applied and when it is applied are shown in tables 6 and 7 respectively. In table 6, it was observed that the performances of the ensemble-based intrusion classification models without feature selection are not too promising. As shown in table 7, Extra Trees model achieved an accuracy of 0.97, precision of 0.98, recall of 0.98 and f1-score of 0.98. Similarly, Random Forest model achieved an accuracy of 0.98, precision of 0.98, recall of 0.99 and f1-score of 0.98. Similarly, Adaboost-based model had an accuracy of 0.96, precision of 0.96, recall of 0.99 and f1-score of 0.98, all in scenario where feature selection was involved. In all, it was deduced that Random Forest intrusion classification model achieved best results when compared to the two other models. Based on the general results in all the homogeneous ensembles that used Decision Trees as base classifiers, the intrusion classification models achieved very promising results in identifying intrusions.

5. Conclusion

This study investigated the performances of three ensemble learning algorithms for the identification of intrusions in UNR-IDD dataset. Decision Trees were used in all the three ensembles. Experimental analyses started from understanding the patterns in the dataset. The insights gained were used to make decision regarding dataset pre-processing. In over all, it can be deduced that Random Forest intrusion classification model achieved the overall best results when compared to the two other models in the intrusion classification task. Based on the general results in all the homogeneous ensembles that use Decision Trees as base classifiers, one can conclude that the models achieved very promising results in the classification of intrusions based on the filter-based attribute selection technique used in the study. The study established that with the right pre-processing of features in the dataset and the use of feature importance as attribute selection technique in Tree-based ensembles, promising classification results were achieved in all the models built from the network data. The study thus argued that tree-based ensembles are generally promising for the identification of intrusions in the chosen dataset. Experimental results further revealed that RF-based intrusion identification model has the overall best performances when compared with the two other intrusion classification models. It is concluded that the three homogeneous ensemble models achieved very promising results while feature importance was used as attribute selection method compared to when no feature selection technique is used.

6. Acknowledgement

The authors acknowledge the feedbacks of the anonymous reviewers which greatly helped in improving the manuscript.

References

[1] R, Tahri, Y. Balouki, A. Jarrar & L. Abdellatif, “Intrusion detection system using machine learning algorithms”, ITM Web of Conferences 46(6):02003, 2022,DOI: 10.1051/itmconf/20224602003

[2] T. Das. T., O. A. Hamdan., R. M. Shukla, S. Sengupta & E. Arslan, “UNR-IDD: Intrusion detection dataset using Network Port Statistics”, *2023 IEEE 20th Consumer Communications & Networking Conference (CCNC)*, Las Vegas, NV, USA, 2023, pp. 497-500, doi: 10.1109/CCNC51644.2023.10059640.

[3] P. Singh P. (2019). “Supervised machine learning”. In: Learn PySpark. Springer. pp. 117–59

[4] Y. Freund, & R. Schapire, “Experiments with a new boosting algorithm”, *In Proceedings of the Thirteenth International Conference on Machine Learning*,1996 pp. 148–156 Bari, Italy.

[5] E. Bauer & R. Kohavi (1999). “Comparison of voting classification algorithms: Bagging, boosting and variants”. Machine An empirical Learning, 36(1/2), 525–536

[6] R. G. Jimoh, A. M. Oyelakin, O. C. Abikoye, M. B. Akanbi, M. D. Gbolagade, A. O. Akanni, M. A. Jibrin & T. S. Ogundele (2023). Efficient Ensemble-based Phishing Website Classification Models using Feature Importance Attribute Selection and Hyper parameter Tuning Approaches, *Journal of Information Technology and Computing,* 2023, 4(2): 1 – 10 DOI : 10.48185/jitc.v4i2.891

[7] A. M. Oyelakin, M. B. Akanbi, T. S. Ogundele, A. O. Akanni ,M. D. Gbolagade, M. D Rilwan, & M. A. Jibrin (2023).”A Machine Learning Approach for the Identification of Network Intrusions Based on Ensemble XGBOOST Classifier”, *Indonesian Journal of Data and Science, 4(3),190-19, 2023*

[8] Y. S. Almutairi,B. Alhazmi & A. A. Munshi, “Network Intrusion Detection Using Machine Learning Techniques”, Advances in Science and Technology Research Journal 2022, 16(3), 193–206, ,https://doi.org/10.12913/22998624/149934

[9] A. M. Oyelakin, M. O. Alimi, I. O. Mustapha & I. K. Ajiboye, “Analysis of Single and Ensemble Machine Learning Classifiers for Phishing Attacks Detection”. *International Journal of Software Engineering and Computer Systems, 7*(2), 44–49, Faculty of Computing, College of Computing and Applied Sciences, Universiti Malaysia Pahang,2021, https://doi.org/10.15282/ijsecs.7.2.2021.5.0088

[10] A. Hadeel, S. Ahmad & E. S. Khair, “A feature selection algorithm for intrusion detection system based on Pigeon Inspired Optimizer”, Expert Systems with Applications, 2020, https://doi.org/10.1016/j.eswa.2020.11324

[11] S. M. Kasongo & Y. Sun ,”Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset”. *Journal of Big Data* **7**, 105, 2020, https://doi.org/10.1186/s40537-020-00379-6

[12] C. Khammassi & S. Krichen (2017). *"GA-LR wrapper approach for feature selection in network intrusion detection",. Computers & Security, 70(), 255–277*. doi:10.1016/j.cose.2017.06.005

[13] L. Breiman "Bagging Predictors, Machine Learning", 26, No. 2, 123-140,2006

[14] G. Pierre, E. Damien & W. Louis," Extremely randomized trees. In Machine Learning", pages 3–42. Machine Learning, 2006

[15] Y. Freund & R. E. Schapire, "A Short Introduction to Boosting, *Journal of Japanese Society for Artificial Intelligence"*, 14(5):771-780, September, 1999. (In Japanese, translation by Naoki Abe, 1-14, https://cseweb.ucsd.edu/~yfreund/papers/IntroToBoosting.pdf

[16] A. M. Oyelakin**,** A. O. Ameen, T. S. Ogundele, T. T. Salau-Ibrahim, U. T. Abdulrauf , \H.I. Olufadi, I. K. Ajiboye, S. Muhammad-Thani , & I. A. Adeniji . "Overview and Exploratory Analyses of CICIDS 2017 Intrusion Detection Dataset", Journal of Systems Engineering and Information Technology (JOSEIT), 2(2), 45-52. https://doi.org/10.29207/joseit.v2i2.5411,2023

[17] S. B. Kotsiantis (2011)."Feature selection for machine learning classification problems: a recent overview",2011

[18] M. Banerjee, E. Reynolds , H. B. Andersson & B. K. Nallamothu" Tree-Based Analysis. Circ Cardiovasc Qual Outcomes."; 12(5):e004879, 2019 doi: 10.1161/CIRCOUTCOMES.118.004879. Erratum in: Circ Cardiovasc Qual Outcomes. 2019 Jun;12(6):e000056. PMID: 31043064; PMCID: PMC6555420.

[19] G. Dong & H. Liu. "Feature engineering for machine learning and data analytics." Boca Raton: CRC Press; 2018.