Journal of Information Technology and Computing
https://jitc.sabapub.com
ISSN: 2709-5916

2023 Volume 4, Issue 2 : 1 – 10
DOI : 10.48185/jitc.v4i2.891

Efficient Ensemble-based Phishing Website Classification Models using Feature Importance Attribute Selection and Hyper parameter Tuning Approaches

Jimoh R. G[1]., Oyelakin A. M. [2,*], Abikoye O. C. [1], Akanbi M. B. [3], Gbolagade M. D. [2], Akanni A. O. [2], Jibrin M. A. [4], Ogundele T. S. [4]

[1]Department of Computer Science, University of Ilorin, Ilorin, Nigeria
[2]Department of Computer Science, Al-Hikmah University, Ilorin, Nigeria
[3]Department of Computer Science, Kwara State Polytechnic, Ilorin, Nigeria
[4]Part Time Lecturer, Department of Computer Science, Al-Hikmah University, Ilorin, Nigeria

Received: 04.11.2023 • Accepted: 15.12.2023 • Published: 30.12.2023 • Final Version: 30.12.2023

Abstract: The internet is now a common place for different business, scientific and educational activities. However, there are bad elements in the internet space that keep using different attack techniques to perpetrate evils. Among these categories are people who use phishing techniques to launch attacks in the enterprise networks and internet space. The use of machine learning (ML) approaches for phishing attacks classification is an active research area in the field of cyber security. This is because phishing attack detection is a good example of intrusion identification tasks. These machine learning techniques can be categorized as single and ensemble learners. Ensemble learners have been identified to be more promising than the single classifiers. However, some of the ways to achieve an improved ML-based detection models are through feature selection/dimensionality reduction as well as hyper parameter tuning. This study focuses on the classification of phishing websites using ensemble learning algorithms. Random Forest (RF) and Extra Trees ensembles were used for the phishing classification. The models built from the algorithms are optimized by applying a feature importance attribute selection and hyper parameter tuning approaches. The RF-based phishing classification model achieved 99.3% accuracy, 0.996 recall, 0.983 f1-score, 0.996 precision and 1.000 as AUC score. Similarly, Extra Trees-based model attained 99.1% accuracy, 0.990 as recall, F1-score was 0.981, precision of 0.990 while AUC score is 1.000. Thus, the RF-based phishing classification model slightly achieved better classification results when compared with the Extra Trees own. The study concluded that attribute selection and hyper parameter tuning approaches employed are very promising.

Keywords: Phishing, Cyber Security, Classification Models, Hyper parameter Tuning

1. Introduction

The internet is a common place for different business, scientific and educational, business activities. The internet broke geographical barriers and allows people to interact, learn, and do businesses together irrespective of their geographical locations. However, there are bad elements with malicious intent that keep using the internet to perpetrate evil. Among these categories are people who use different spam and phishing techniques to launch attacks in the internet (Adewale, & Olugbara, 2017). Oyelakin (2014) mentioned that there is growing cases of spear phishing attacks in the internet space and described how spam attackers are using phishing to harvest the sensitive credentials of

* Corresponding Author: amoyelakin@alhikmah.edu.ng

unsuspecting bank account holders. The study reported statistical evidence of how online awareness among bank account holders in Nigeria can be of great help to stem the negative trends.

Signature and ML-based techniques are widely used for phishing classification and related cyber security attacks. However, Pektas et al. (2018) has argued that the use of these approaches for the classification of different types of intrusions attacks is getting popular compared to signature-based methods. Specifically, other researchers have re-echoed how the use of supervised machine learning techniques have been very renowned for phishing attack classification in recent times (Li et al., 2019; Oyelakin, Alimi & Abdulrauf., 2020; Oyelakin, Olatinwo., Rilwan., Azeez & Obiwusi 2021a; Mohanty & Acharya, 2023). Li et al., (2019;) and Oyelakin et al. (2020). Have pointed out that some of the supervised learning algorithms that have been used for security tasks include are Naïve Bayes, Logistic Regression, decision trees, Support Vector Machines and ensemble learners.

Figure 1. Website Phishing Representation (Martin, 2022)

In the fourth quarter of year 2022, APWG reported a total of one million, three hundred and fifty thousand and thirty seven (1,350,037) phishing attacks. APWG (2022) further argued that the figure was up slightly from the third quarter of the same year when APWG claimed that there were 1,270,883 cases of phishing. Bad actors in the internet space used different ways to launch phishing attacks. For instance, the threat actors in phishing attacks may try to present themselves as colleagues, acquaintances, reputable organizations and then solicit sensitive information or try to lure victims into downloading files which may execute as malware (Mohammed et al., 2014).

Phishing is the art of emulating a website of a creditable firm intending to grab user's private information such as usernames, passwords and social security number (Mohammed et al., 2014). Ensemble learning methods are made up of a set of classifiers such as decision trees and their predictions are aggregated to identify the most popular classification result. Examples of ensemble methods include Random Forest, Extra Trees, AdaBoost, XGBoost and many others. These algorithms build many trees in the process. In the end, the final prediction is based on all of the trees.

Aside, Jimoh, Oyelakin, Olatinwo , Obiwusi, Muhammad-Thani, Ogundele, Giwa-Raheem and Ayepeku (2022) have mentioned that ensemble learning approaches are promising for spam classification on Twitter platform. Aside this**,** Yang and Shami (2022) argued that selecting the best hyper-parameter configuration for machine learning models directly affects their performances. This study aims at applying RandomSearch approach for the hyper parameter tuning of the learning algorithms while feature importance is used for the feature subset selection. Thereafter, Random forest and ExtraTrees ensemble learners are used for the identification of phishing attacks in this study. Random Forest (RF) was put forward by Breiman (2001). ExtraTrees was originally proposed by Pierre, Damien and· Louis in 2006. It is a tree-based ensemble method for supervised classification and regression problems (Pierre, Damien & Louis, 2006). The study focuses on

extending previous works by Oyelakin et al. (2021a). This study focuses on investigating how improvement can be achieved in phishing website classification based on the use of feature importance for feature selection and hyper parameter tuning for optimising the phishing classification model performances.

2. Related works

Aljammal, Taamneh, Qawasmeh and Bani (2023) built six machine learning models using variety of classifiers. The selected algorithms were trained and tested using phishing datasets both with and without feature selection. Authors argued that out of the algorithms, Random Forest classifier was superior in performance as it achieved accuracy of 98% and 93.66% respectively for the chosen datasets. Mohanty and Acharya (2023) proposed a detection framework for identifying suspicious web sites with the help of a multivariate filter-based feature selection technique. A correlation feature selection approach was employed. Lastly, three different ensembles and kNN classifiers were used for the prediction of the malicious web sites efficiently. The authors evaluated the classifier with and without considering the attribute selection. He further mentioned that the implementation results are promising as the learning algorithms accomplished the highest classification accuracy of 97% in dataset I and 99.25% in the second dataset based on the attribute selection method used.

Similarly, Oyelakin et al. (2021a) carried out an investigation into the performances of supervised learning algorithms for the identification of phishing attacks by applying different phishing datasets. A filter-based feature selection method called ANOVA F-test was used to select promising features. Then, four classification models were built. Authors argued that Random Forest algorithm has the best performances based on the selected metrics. Oyelakin, Alimi, Mustapha and Ajiboye (2021b) built single and ensemble learning models for phishing attacks classification. It was argued that RF method was very promising compared to others. Similarly, Oyelakin, Alimi and Abdulrauf (2020) used some learning algorithms to build phishing URL classification models. The study reported promising results and argued that ML techniques are better than traditional methods in phishing identification problems.

Moreover, Hossain, Sarma and Chakma (2020) used machine learning techniques to build phishing detection models and evaluated their performances. The study used algorithms like KNN, SGD, and Random Forest as the learning algorithms for building the models. It was argued that Random Forest classifier performed better across the chosen metrics. Apart from this, Oyelakin et al. (2020) compared how some selected ML Algorithms behave in the classification of Phishing URLs. The9 study contributed to the development of this project by informing the selection and evaluation of machine learning techniques for spam URL classification. Patil and Patil (2018) used supervised decision tree learning classification algorithms to build models. They performed experiments on the balanced dataset. Authors argued that they achieved experimental results which showed 99.29% detection accuracy.

Orji and Emekwuru (2019) compared selected ML algorithms for phishing website classification. The authors evaluated five different algorithms in the chosen phishing dataset. They reported that RF and SVM models achieved the highest accuracy and precision. Apart from this, Biswas et al. (2018) investigated various feature engineering and selection techniques for spam URL classification. The authors examined different URL attributes, such as domain reputation, URL length, and presence of specific keywords, and evaluated their impact on classification performance. Although Extra Trees was not explicitly used in this study, the findings regarding feature engineering

and selection strategies provided insights that can be applied when utilizing Extra Trees for spam URL classification.

3. Methodology

3.1. Problem Description

The problem at hand is a supervised binary classification one. It involves building two different ensemble-based learners for the classification of phishing evidence. The target is to achieve models that have the ability to efficiently classify the dataset used for the experimentation as phishing and non-phishing promising results across the five selected metrics. The two algorithms used are all tree-based ensembles. Feature importance attribute selection and Grid Search hyper parameter tuning techniques were used so as to optimize the proposed model performances. The feature importance was used for the attribute selection while Random search was employed for the tuning in this study. Yang et al. (2022) established that hyper parameter tuning is very promising in ML researches.. The hyper parameter values were set before the training process. It was argued that checks a randomly selected fixed number of combinations specified in n_iter of the RandomizedSearchCV function. Random search has a very high probability of finding the optimal hyper parameter combination within the randomly selected combinations. Hyper parameter optimization was carried out in the experiments for the two ML-based phishing classification models.

3.2. Dataset collection and Description

The dataset used in this study was collected from UCI Machine Learning Repository. The dataset was released by Mohammad, McCluskey and Thabtah (2014). Basic characteristics are shown in table 1.

Table 1. Dataset Characteristics

No of Attributes	No of Instances	Any Missing values?	Are the Data Types Mixed?
30	11054	No	No. The input attributes are numeric while the target class is categorical.

The dataset consists of a collection of website URLs for 11054 websites. Each sample has 30 website attributes and a class label identifying it as a phishing website or not (1 or -1). Some of the attributes/features in the dataset include Index,UsingIP, LongURL, ShortURL,Symbol@, Redirecting//, PrefixSuffix-, SubDomains, HTTPS, DomainRegLen and so on.

3.3. Data Preprocessing

The dataset used for the study consists of input features that are numeric in nature while the target attribute is categorical. The only data pre-processing step taken is to scale the features so that the learning algorithms will not be biased towards the phishing classification task.

3.4. Model Development

The dataset was split into the train test ratio of 80 to 20. A combination of hyper parameter settings were used for the model building. Random Forest and Extra Tree models were fitted. The best hyperparamters are used for the model performance tuning in each of the scenarios. For the attribute selection, feature importance was used in the Tree-based ensemble learners. Figure 1 is used to pictorially represent the various processes through which the classification of phishing attacks in the chosen dataset was arrived at.

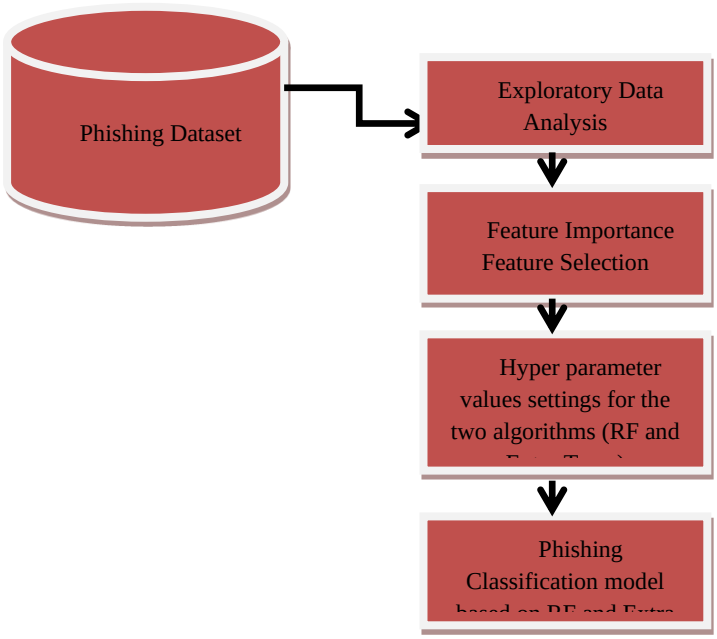

Figure 2. Methodological Process in the Study

The values for hyper parameters were set at the creation of the RF and Extra Tree model. The feature scores obtained based from the feature importance technique were visualized for the two selected algorithms. The performances of the models were then evaluated using the identified metrics: accuracy, recall, f1-score, precision and Area Under the Curve (AUC).

Algorithm 1: Algorithm for Random Forest Phishing Classification

Input-Given a phishing website dataset with some set of features as inputs

Output: results achieved by RF classifier based on Accuracy, precision and other metrics selected

Pick random samples from a given data or training set.

Construct a decision tree for every training data

Compute the voting by averaging the decision tree.

Finally, pick the most voted classification result as the final result based on the Decision Trees used.

Output the classification results

Algorithm 2: Algorithm for Extra Trees for Phishing Classification

Split a node(S)

Input: Given a phishing website dataset with some set of features as inputs to the node we want to split

Output: a split [a < ac] or nothing

If Stop split(S) is TRUE then return nothing.

Otherwise select K attributes from the phishing dataset {a1,..., aK } among all non constant (in S) candidate attributes;

Draw K splits {s1,...,sK }, where si = Pick a random split(S, ai), ∀i = 1,..., K;

Return a split s∗ such that Score (s∗, S) = maxi=1,...,K Score(si, S).

Pick a random split from the phishing website dataset(S,a)

Select the most promising classification result based on the splitting

Output the classification results

4. RESULTS AND DISCUSSION

4.1. Results

Results of selected attributes

In the Tree-based Random Forest ensemble, fifteen (15) features with promising scores were selected based on the threshold set. A threshold of 0.01 was set to arrive at the selected features for building the model. The features are as visualized as shown in figure 1.

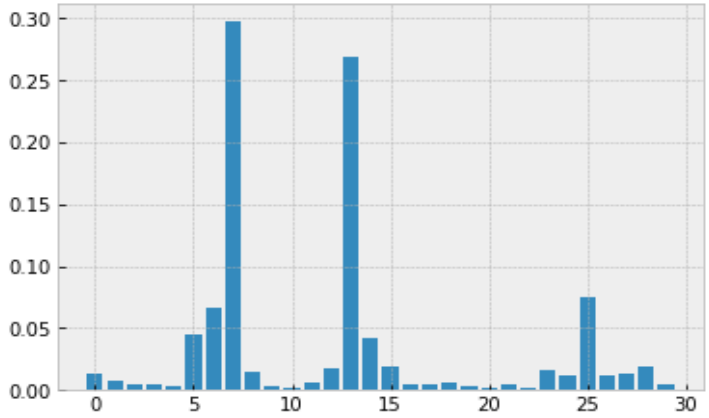

Figure 3. Feature Importance Scores for Random Forest Model

In the tree-based Extra Tree ensemble, eleven (11) features with promising scores were selected based on the threshold set. A threshold of 0.01 was set to arrive at the selected features for building the model. The features are as visualized as shown in figure 3.

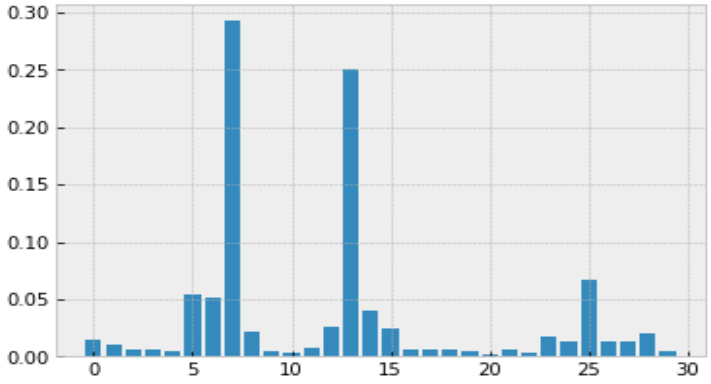

Figure 4. Feature Importance Scores for Extra Trees Model

Table 2. Results of RF-based Phishing Classification Model

RF Model	Results of RF-based Phishing Classification using Hyperparameter Tuning
Accuracy (%)	99.30
Recall	0.996
F1-Score	0.983
Precision	0.996
AUC Score	1.000

The results of the phishing website classification based on the identified promising features are as shown in table 2.

AUC Score visualization for RF-based model

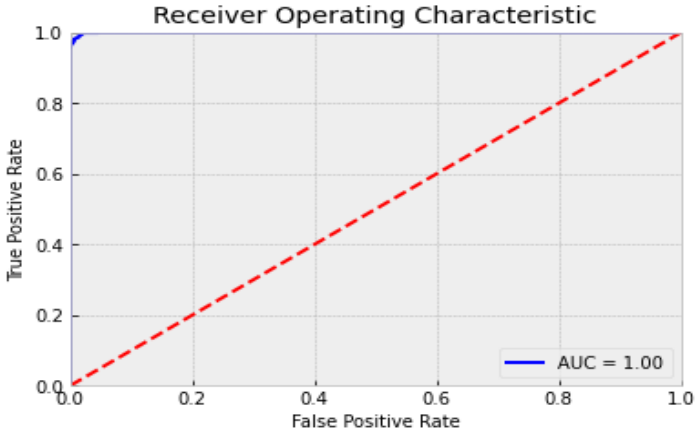

Figure 5. AUC-ROC score visualisation for RF-based Model

Table 3. Table Results of Extra Trees-based Phishing Classification Model

Extra Trees Model	Results of Extra Trees-based Phishing Classification using Hyperparameter Tuning
Accuracy (%)	99.10
Recall	0.990
F1-Score	0.981
Precision	0.990
AUC Score	1.000

The results of the phishing website classification based on the identified promising features are as shown in table 1.

AUC Score Visualisation for Extra Tree-based Model

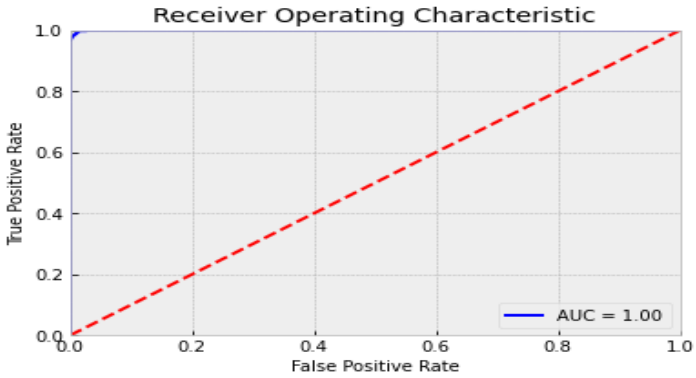

Figure 6. AUC-ROC score visualisation for Extra Trees Model

4.2. Discussion of Results

First of all, exploratory analysis was carried out on the chosen dataset. The analysis of the dataset revealed the basic characteristics of the data. This informed the choice of the feature selection approach used. The two ML-based models (RF and Extra Trees models) were able to achieve enhanced results owing to their ability to efficiently classify the dataset used for the experimentation as phishing and non-phishing promising results across the five selected metrics. The two algorithms (RF and Extra Trees) used are all tree-based ensembles and were applied for building the phishing classification models. Feature importance attribute selection and Grid Search hyper parameter tuning techniques were used to achieve the improvement. The RF-based model achieved 99.3% accuracy, 0.996 recall, 0.983 f1-score, 0.996 precision and 1.000 as AUC score. Similarly, Extra Trees-based model attained 99.1% accuracy, 0.990 as recall, F1-score was 0.981, precision of 0.990 while AUC score is 1.000. Thus, the RF-based phishing classification model slightly achieved better classification results when compared with the Extra Trees model. This study was also benchmarked with two similar studies that used the same phishing dataset in recent years. It was shown that the results achieved by the two ensemble approaches used in this paper are better. Thus, this study has demonstrated the effect of feature selection and optimization of machine learning-based models in the classification of phishing attacks.

Benchmarking of the results with similar studies

This study was benchmarked with two similar studies that used the same phishing dataset in recent years. The two ML-based approaches were able to achieve enhanced models that have the ability to efficiently classify the dataset used for the experimentation as phishing and non-phishing promising results across the five selected metrics. The two algorithms (RF and Extra Trees) used are all tree-based ensembles. It is evident that the results obtained in the two models are slightly better than the ones in similar studies by Oyelakin et al. (2021) and Hossain et al.(2020).

Conclusion

This study introduced phishing attacks as one of the key problems confronting the internet community globally. The work also pointed out that ML approaches have been found to be very prominent for handling security related classification or regression problems. The study collected phishing website and performed exploratory analysis of the dataset with a view to understanding the features and instances therein. A filter-based attribute selection method named Feature importance attribute selection was used. Then, Grid Search hyper parameter tuning technique was employed for

the optimisation. The two models built achieved greater performance with the use of the approaches. Experimental results showed that the RF-based model slightly achieved better classification results when compared with the Extra Trees-based model. This paper demonstrated the strengths of feature selection and optimization of ML algorithms in ML-based phishing identification models. This study concluded that attribute selection and hyper parameter tuning approaches employed are very promising.

References

[1] Adewale, O. S., & Olugbara, O. O. (2017). A Comparative Study of Machine Learning Algorithms for Email Spam Filtering, *Expert Systems with Applications*, 74, 219-236.

[2] Aljammal, A. H., Taamneh , S. ., Qawasmeh, A. ., & Bani Salameh, H. (2023). Machine Learning Based Phishing Attacks Detection Using Multiple Datasets. *International Journal of Interactive Mobile Technologies (iJIM)*, *17*(05), pp. 71–83. https://doi.org/10.3991/ijim.v17i05.37575

[3] APWG (2022). Phishing Activity Trends Report, 4th Quarter 2022, Unifying the Global Response To Cybercrime, Activity October - December 2022, https://docs.apwg.org/reports/apwg_trends_report_q4_2022.pdf

[4] Biswas, A., Dasgupta, A., & Nag, P. K. (2018). Feature Engineering and Selection for Spam URL Classification, *International Journal of Computer Applications*, 179(30), 25-28.

[5] Breiman L. (2001). Random Forests, Machine Learning, 45(1), 5-32, (2001). Available at: https://doi.org/10.1023/A:1010933404324

[6] Hossain Sohrab, Sarma Dhiman & Chakma R. (2020). Machine Learning-Based Phishing Attack Detection, International Journal of Advanced Computer Science and Applications (IJACSA), (11)9, 2020DOI:10.14569/ijacsa.2020.0110945Corpus ID: 222469828

[7] Jimoh R. G., Oyelakin A. M. Olatinwo , I. S., Obiwusi Y. K., Muhammad-Thani S., Ogundele T. S., Giwa-Raheem A. & Ayepeku O. F. (2022). Experimental Evaluation of Ensemble Learning-Based Models for Twitter Spam Classification, *2022 5th Information Technology for Education and Development (ITED) conference, held at Nile University Abuja, Nigeria*

[8] Li, X., & Li, X. (2019). Web page classification using machine learning: A comprehensive survey. *ACM Computing Surveys*, 52(6), 1-34.

[9] Mohammad,Rami and McCluskey,Lee. (2015). Phishing Websites. UCI Machine Learning Repository. https://doi.org/10.24432/C51W2X

[10] Martin Jessica (2022). How phishing can ruin the good name of an online brand, published by *reputation*, retrieved from https://blog.reputationx.com/guest/whats-phishing on 1st July, 2023

[11] Mohammad, *Rami M.*, Thabtah, *Fadi* & McCluskey, Lee. (2014). Intelligent Rule based Phishing Websites Classification. IET Information Security, 8 (3), 153-160. 2014, 1751-8709, available at https://archive.ics.uci.edu/ml/machine-learning-databases/00327/

[12] Mohanty Sanjukta & Acharya Arup Abhinna (2023). MFBFST: Building a stable ensemble learning model using multivariate filter-based feature selection technique for detection of suspicious URL, *Procedia Computer Science*, Volume 218, 2023, Pages 1668-1681

[13] Orji, I. J., & Emekwuru, O. E. (2019). Comparative Analysis of Machine Learning Algorithms for Phishing Website Detection. *International Journal of Computer Science and Information Technology Research*, 7(2), 98-106.

[14] Oyelakin A. M., Olatinwo I. S., Rilwan D. M., Azeez R. D. & Obiwusi Y. K (2021a). Investigation into the Performances of Supervised Learning Algorithms in different Phishing Datasets, *Pakistan Journal of Engineering Technology and Science (PJETS)*, 9(2), 24-32

[15] Oyelakin A. M., Alimi M. O., Mustapha I.O. & Ajiboye I. K. (2021b). 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, https://doi.org/10.15282/ijsecs.7.2.2021.5.0088

[16] Oyelakin A. M., Alimi O. M., & Abdulrauf T. (2020). Performance Analysis of Selected Machine Learning Algorithms for the Classification of Phishing URLs, *Journal of Computer Science and Control Systems*, 13(2), 16–19 , available at

https://electroinf.uoradea.ro/images/articles/CERCETARE/Reviste/JCSCS/JCSC_V13_N2_oct2020/JCSCS VOL 13 NO 2 OCTOBER 2020 Oyelakin_Performance.pdf

[17] Oyelakin A. M. (2014). Spear Phishing Email Attack on Nigerian Bank Account Holders: Online Awareness to the Rescue**,** *in the proceedings of ISTEAM Conference 2014, Afe Babalola University, Ado Ekiti, Nigeria*, 185-188

[18] Patil Dharmaraj R. & Patil Jayantrao (2018). Malicious URLs Detection Using Decision Tree Classifiers and Majority Voting Technique, *Cybernetics and Information Technologies* 18(1):11-29, DOI: , 10.2478/cait-2018-0002

[19] Pierre Geurts, Damien Ernst & Louis Wehenkel (2006). Extremely randomized trees, Machine Learning, 63: 3–42, DOI:10.1007/s10994-006-6226-1https://link.springer.com/content/pdf/10.1007/s10994-006-6226-1.pdf

[20] Yang Li and Shami Abdallah (2022).On Hyperparameter Optimization of Machine Learning Algorithms: Theory and Practice, a preprint retrieved from arXiv:2007.15745v3 [cs.LG] 5 Oct 2022