INTERNATIONAL JOURNAL OF SOFTWARE ENGINEERING & COMPUTER SYSTEMS (IJSECS)
ISSN:2289-8522 e-ISSN:2180-0650
VOL. 7, ISSUE 2,44 –49
DOI: https://doi.org/10.15282/ijsecs.7.2.2021.5.0088

ORIGINAL ARTICLE

Analysis of Single and Ensemble Machine Learning Classifiers for Phishing Attacks Detection

[1]Oyelakin A.M., [1]Alimi O. M., [1]Mustapha I. O, [2]Ajiboye I. K.
[1]Department of Computer Science, Faculty of Natural and Applied Sciences
Al-Hikmah University, Ilorin, Nigeria
[2]Abdulraheem College of Advanced Studies, Igbaja

ABSTRACT–Phishing attacks have been used in different ways to harvest the confidential information of unsuspecting internet users. To stem the tide of phishing-based attacks, several machine learning techniques have been proposed in the past. However, fewer studies have considered investigating single and ensemble machine learning-based models for the classification of phishing attacks. This study carried out performance analysis of selected single and ensemble machine learning (ML) classifiers in phishing classification.The focus is to investigate how these algorithms behave in the classification of phishing attacks in the chosen dataset. Logistic Regression and Decision Trees were chosen as single learning classifiers while simple voting techniques and Random Forest were used as the ensemble machine learning algorithms. Accuracy, Precision, Recall and F1-score were used as performance metrics. Logistic Regression algorithm recorded 0.86 as accuracy, 0.89 as precision, 0.87 as recall and 0.81 as F1-score. Similarly, the Decision Trees classifier achieved an accuracy of 0.87, 0.83 for precision, 0.88 for recall and 0.81 for F1-score. In the voting ensemble, accuracy of 0.92 was achieved. 0.90 was obtained for precision, 0.92 for recall and 0.92 for F1-score. Random Forest algorithm recorded 0.98, 0.97, 0.98 and 0.97 as accuracy, precision, recall and F1-score respectively. From the experimental analyses, Random Forest algorithm outperformed simple averaging classifier and the two single algorithms used for phishing url detection. The study established that the ensemble techniques that were used for the experimentations are more efficient for phishing url identification compared to the single classifiers.

ARTICLE HISTORY
Received:15 May 2021
Revised:30 Aug 2021
Accepted:17 Sept 2021

KEYWORDS
Phishing Attacks, Internet Security, Ensemble Machine Learning Algorithms, Classification

INTRODUCTION

Different methods have been used by attackers to launch phishing-based attacks in networks and the internet space. Specifically, these phishing techniques are used by cyber criminals for stealing online users' personal identity as well as financial account credentials [1]. Authors in [1] reported that for the first quarter of 2020 revealed that phishing attacks have risen greatly beyond the past years. The report had it that 60,000 of phishing sites were reported in March 2020 alone. To counter the different phishing-based attacks, researchers have been employing signature-based and machine learning approaches. However, machine learning approaches have been found more suitable for phishing detection compared to signature-based techniques [2]. As part of the efforts to solve the problem of phishing attacks. [2] further argued that there are three major techniques for phishing detection. He mentioned context based technique, URL based method and machine learning technique.

The problem identified with signature-based phishing detection approaches makes machine learning techniques to be getting popular in phishing detection studies ([3]; [4]). The machine learning techniques can be classified into four. They are: supervised, unsupervised, semi-supervised learning and reinforcement learning techniques [5]. In supervised learning algorithms, labeled datasets are provided, and the algorithm uses them for the training and testing. As reported in different literature, some examples of supervised machine learning algorithms that can be used for classification or regression tasks include: Logistic Regression, Naïve Bayes, Decision Tree Classifier, Support Vector Machine, K Nearest Neighbour, and ensemble algorithms. Good examples of ensemble classifiers are ExtraTree algorithm, Adaboost, Random Forest Algorithm, Voting Classifier, XGBoost and so on. It has been argued that the ensemble classifiers are generally more accurate than any of the individual classifiers ([6]; [7]; [8]; [9];[10]).

The phishing detection problem in this study is handled using four supervised machine learning algorithms. The classifiers are single and ensemble types. In general, every supervised learning algorithm consists of a target variable which is to be predicted from a given set of predictors [5]. This paper aims at investigating the performances of selected single and ensemble learning algorithms in the detection of phishing-based attacks. The study builds phishing detection models from the algorithms. Generally, single and ensemble algorithms behave differently and this serves as the justification to investigate how the two categories of the algorithms behave in phishing classification. While the single classifier is regarded as weak learners, the ensembles are built from individual performances of the weak classifiers. Then,

*CORRESPONDING AUTHOR |A. M. Oyelakin | amoyelakin@alhikmah.edu.ng

a comparative analysis of the selected single and ensemble models were carried out with the use of accuracy, precision, recall and F1-score as metrics. The chosen algorithms are popular and representative as types of learning classifiers under the two categories considered in the study.

The rest of this paper is organised as follows. Section two provided review of related studies in phishing detection. The third section described the methodology used for the various stages of machine learning-based phishing detection. Thereafter, the results obtained from the analysis were presented and discussed in the fourth section. These four sections were followed by acknowledgment, conclusion and references.

RELATED WORK

In [11], the researchers proposed machine learning-based models for the prediction of phishing based attacks. The algorithms used for building the models include: Logistic Regression, Support Vector Machines (SVM), Decision Trees and Neural Networks. The authors used a phishing dataset collected from UCI Machine Learning repository. Accuracy, sensitivity and specificity were used as metrics. The paper reported that Support Vector Machine has the largest performance by achieving 89.84% of accuracy, 93% of specificity and 89% of sensitivity.

Similarly, researchers in [12] provided promising evidence for the use of machine learning techniques for botnet detection. The paper provided empirical results from the proposed machine learning methods used for phishing attacks classification. The performances of the selected machine learning algorithms were compared. In the study by [13], the authors carried out a comprehensive review of literature on phishing attack detection. The study provided anti-phishing training and awareness for online users with a view to stemming the tide of phishing based attacks. However, the work only focused on comparative literature review without the need to develop anti phishing solutions.

Apart from this, authors in [4] used four single machine learning classifiers for the identification of phishing evidence in the chosen phishing url dataset. The authors carried out performance analysis of the learning algorithms using a UCI Machine learning repository phishing dataset that was released in 2018. The performance analyses of the algorithms were measured using accuracy, precision, recall and f1-score.However, the work did not consider ensemble learning approach for the phishing classification. Authors in [2] built a model for the classification of phishing attacks. The authors focused on the evaluation of the chosen classifier using only accuracy at the detriment of other useful metrics. Thus, erroneous judgments were arrived at in the phishing classification.

Similarly, [3] used five different machine learning algorithms for the classification of malicious urls. The authors used only three performance metrics for the evaluation of the selected algorithms. Unlike the approach in this study, emphasis was not on comparison of the performances of single and ensemble classifiers.

METHODOLOGY

(i) Dataset and its collection process

The chosen dataset was released publicly by [14], the dataset is publicly available for download in the UCI Machine Learning repository at https://archive.ics.uci.edu/ml/machine-learning-databases/00327/. This dataset was originally collected by the above authors from PhishTank archive, MillerSmiles archive, and Google searching operators. The dataset is suitable for the study since it is current and contains features that are proven to be relevant for identifying phishing attacks. The dataset is made up of numeric features as inputs and categorical features and target. The dataset is originally available in arff format and was converted to csv format. The dataset in its csv format was exported into the Python environment where different experimental analyses were carried out. The dataset characteristics are as shown in Table 1.

Table 1. Dataset Characteristics

S/N	Characteristics	Value
1	Missing values	No
2	Input variables	Numeric
3	Target variable	Categorical
4	No of instances (samples)	11054
5	No of features(input variables)	31

(ii) Data Preprocessing and Feature Selection Method Used

Exploratory Data Analysis (EDA) was carried out on the dataset. The analysis showed the basic characteristics of the dataset. Then, the dataset was split in the ratio 80:20 as training and testing sets respectively. This is to enable us to train the selected algorithms on the training set, and make predictions on the test set respectively. As part of the pre-processing, the dataset is scaled so as to achieve improved predictive ability. Also, ANOVA F-test was used as a feature-selected

technique. In each of the algorithms used for the experimentations in this study, features that are independent of the target variable are removed from the dataset. The choice of the selection method is based on the nature of the data values in the dataset. There is a need for feature selection in a machine learning-based classification task so as to improve the predictive ability, reduce training time and improve interpretability [5].

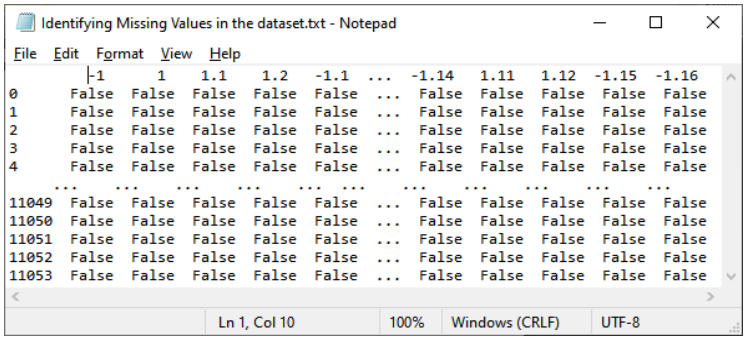

Figure 1. EDA showing no missing values

Figure 1 provides exploratory data analysis whether there are missing values in the dataset or not. Exploratory data analysis shows that there are no missing values.

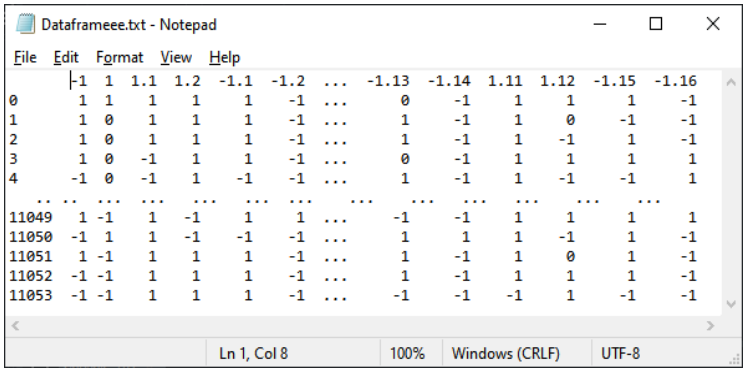

Figure 2. EDA showing the information in the dataframe

Figure 2 provides information in the dataset through exploratory analysis carried out to determine if there are missing values in the dataset. No missing values are found.

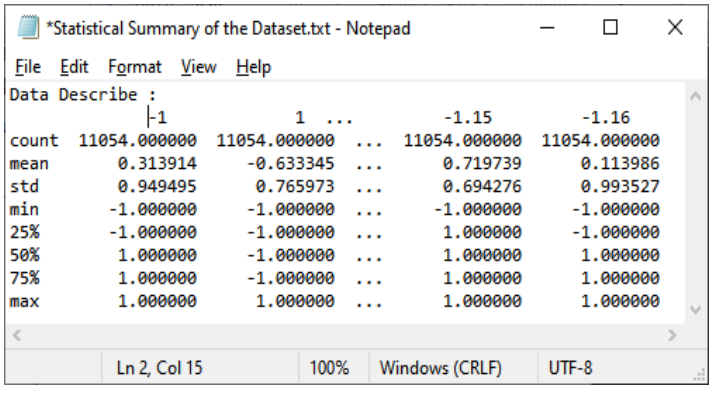

Figure 3. Statistical summary of the features in the dataset

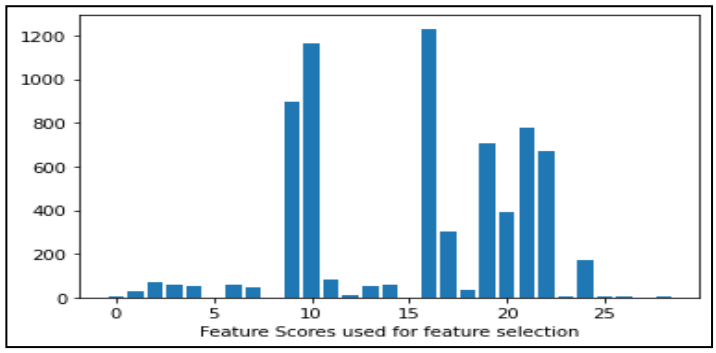

Figure 4. ANOVA F-scores for the dataset

Figure 4 is used to represent the scores of various features in the dataset. Depending on the score assigned a particular feature; the attribute is selected or dropped. In this work, features with the larger scores were selected while the ones with low scores are dropped.

Phishing Classification using selected algorithms

The algorithms used for the classification fall into single and ensemble groups. The ways each of the two algorithms behave are briefly described below.

Single ML Classification Algorithms

The single machine learning algorithms considered for empirical analysis in this study are: Logistic Regression and Decision Tree. Both Logistic regression and Decision Tree classifiers are supervised machine learning algorithms that are used for prediction or regression tasks [5]. The two algorithms support binary classification problems in the phishing detection under consideration.

Ensemble ML Classification Algorithms

Ensemble models in machine learning operate by combining the decisions from multiple single models to improve the overall performance. This is one of the reasons why they are termed strong classifiers as against the single one that are called weak learners. Examples of ensemble classifiers are simple voting methods, Random Forest, AdaBoost and stacking algorithms. Random Forest is an 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 robustness of the weak learner. As reported in literature, the main reason behind ensemble is to build a more robust classifier from the weaker ones ([3]; [9]). The ensemble methods used in this study include: Simple Averaging technique and a Bagging technique named Random Forest Classifier.

At the phishing classification stage, the pre-processed dataset was fed into each of the learning algorithms at different times and phishing evidence was classified. The results of the classification were obtained and tabulated as shown in Table 2.

RESULTS AND DISCUSSION

All the experiments were run in the Python 3.7 environment. The tool was chosen because of its robust libraries and full support for developing machine learning based models [17]. The dataset was normalised by scaling its features. Also, the Confusion matrix was used as part of the evaluation process in the work. The confusion matrix contains the individual values of True Positive (TP), True Negative (TN), False Negative (FN), and False Positive (FP). The mathematical formulae used for obtaining the values of the performance metrics are shown in equations 1, 2, 3 and 4 respectively:

$$\text{Accuracy} = (TP+TN)/(TP+TN+FP+FN) \quad (1)$$

$$\text{Precision} = TP/(TP+FP) \quad (2)$$

$$\text{Recall} = TP/(TP+FN) \quad (3)$$

$$\text{F1-score} = 2\times (\text{Precision X Recall})/(\text{Precision} + \text{Recall}) \quad (4)$$

Table 2. Performances of the selected ML Algorithms

S/N	ML Algorithm	Accuracy	Precision	Recall	F1-Score
1	Logistic Regression	0.86	0.88	0.86	0.79
2	Decision Trees	0.87	0.83	0.87	0.81

3	Simple voting Ensemble	0.91	0.90	0.92	0.92
4	Random Forest	0.98	0.97	0.98	0.97

Discussions of Results

In this study, the exploration of the dataset showed that its input features are in numeric form while the target feature is in categorical form. The input values were normalised using min-max scaling. Thereafter, four machine learning algorithms were used for the experimental analyses in the Python 3.7 environment. A phishing dataset was selected and the dataset was split in the ratio 80:20 as training and testing sets respectively. This was carried out to enable us to train the selected algorithms on the training set, and make predictions on the test set respectively. The performances of the algorithms in the classification of phishing urls were obtained and used in Table 2. Also, the performances of the classifiers were compared. The first two algorithms are single classifiers while the remaining two are of the ensemble category. The result of the experimental analyses in Table 2 showed that the Random Forest classifier had the overall best performance, followed by the simple voting ensemble method. Decision tree performed slightly better than the Logistic Regression algorithm while considering accuracy as metrics. Moreover, Logistic Regression classifiers performed better than Decision Trees when precision is measured. However, it was observed that Decision Trees had better performances than Logistic Regression algorithms under recall and f1-score metrics respectively. Generally, the study provided empirical evidence that ensemble machine learning techniques performed better than their single learning classifiers in phishing detection.

CONCLUSION

This study provided a general introduction to machine learning-based phishing attack detection. Unlike the approaches used in some of the past studies, two categories of learning classifiers were used to build phishing detection models. Specifically two single and two ensemble machine learning algorithms were used to build models for the identification of phishing-based cyber attacks. The learning algorithms chosen were trained and tested using the phishing dataset. The metrics used for the evaluation were accuracy, precision, recall and F1-score. It was observed that the two ensembles generally performed better than the two selected classifiers. The results of the experimentation showed that the ensemble techniques that were used performed better than the single classifiers in phishing url identification.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers for their comments which helped in improving the article.

REFERENCES

[1] APWG (2020). Phishing Activity Trends Report for Q1 2020 retrieved from https://docs.apwg.org/reports/apwg_trends_report_q1_2020.pdf

[2] A. Alswailem, B. Alabdullah , N. Alrumayh, and A. Alsedrani. Detecting phishing websites using machine learning. *In 2019 2nd International Conference on Computer Applications Information Security* (ICCAIS), 1–6, (2019)

[3] Akshay Sushena Manjeri, Kaushik R., MNV Ajay, C. Nair Priyanka. A Machine Learning Approach for Detecting Malicious Websites using URL Features. *2019 3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 555–561,(2019). https://doi.org/10.1109/iceca.2019.8821879

[4] A. M. Oyelakin , O. M. Alimi, Tosho Abdulrauf. A Comparative Analysis of Machine Learning Algorithms for Detecting Phishing Urls, *Journal of Computer Science and Control Systems, Oredia University, Romania, 13(2):16-19*, (2020) available at https://electroinf.uoradea.ro/index.php/jcscs/12-cercetare/reviste/jcscs/213-1st-issue-vol-13-nr-2.html

[5] M. A. Hall. Correlation-based Feature Selection for Machine Learning, (1999) *PhD Thesis at University of Waikato*

[6] A. Chaudhary, S. Kolhe, & R. Kamal. An improved Random Forest Classifier for multi-class classification. *Information Processing in Agriculture*, (September 2016). https://doi.org/10.1016/j.inpa.2016.08.002

[7] M. Zakariah. Classification of large datasets using Random Forest Algorithm in various applications : Survey. *International Journal of Engineering and Innovative Technology (IJEIT)*, *4*(3), 189–198. (2014)

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

[9] L. Breiman. Stacked regressions. Machine Learning, 24(1), 49–64. (1996)

[10] 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.*

[11] Sagar Patil, Yogesh Shetye, Nilesh Shendage. Detecting Phishing Websites Using Machine Learning, *International Research Journal of Engineering and Technology (IRJET)*,7(2) (2020)

[12] V. Shahrivari, M. D. Muhammad and I. Muhammad. Phishing Detection Using Machine Learning Techniques, available at https://arxiv.org/pdf/2009.11116.pdf (2020)

[13] D. Jampen, G. Gür, T. Sutter & B. Tellenbach. Don ' t click : towards an effective anti - phishing training . A comparative literature review. In *Human-centric Computing and Information Sciences* (2020). https://doi.org/10.1186/s13673-020-00237-7

[14] Rami Mohammad, T.L. McCluskey and Fadi Abdeljaber Thabtah. Intelligent Rule based Phishing Websites Classification. *IET Information Security*, 8 (3), 153-160. ISSN 1751-8709, (2014). Available at https://archive.ics.uci.edu/ml/machine-learning-databases/00327/

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

[17] M. E. Fenner. Machine Learning with Python for Everyone (2020). Free Sample Chapter, Addison Wesley Data and Analytics Series