INVESTIGATION INTO THE PERFORMANCES OF SUPERVISED LEARNING ALGORITHMS IN DIFFERENT PHISHING DATASETS

[1]Oyelakin A. M., [2]Olatinwo I. S.,[3]Rilwan D. M.,[4]Azeez R. D., [5]Obiwusi Y. K.
[1]Department of Computer Science, Faculty of Natural and Applied Sciences
Al-Hikmah University, Ilorin, Nigeria
[2]Department of Computer Science, Federal Polytechnic, Offa, Nigeria
[3]Department of Computer Science, Sheu Idris College of Health Sciences and Technology, Makarfi, Nigeria
[4]ICT Unit, Al-Hikmah University, Ilorin, Nigeria
[5] Department of Mathematics and Computer Science, Summit University, Offa, Nigeria

***Abstract*- Phishing techniques are employed by attackers to obtain the sensitive and confidential information of unsuspecting internet users. To stem the tide of phishing-based attacks in the cyber space, different machine learning techniques have been proposed as better alternatives to the signature-based approaches. This study used a different approach in the detection of phishing evidence in three phishing datasets. The focus of the study is to investigate the performances of Random Forest, K-Nearest Neighbour and Extra Tree algorithms in three different phishing datasets. The first two datasets of different sizes and features were obtained from Machine Learning UCI repository while the third dataset was collected from Mendeley. Exploratory Data Analysis was carried out in order to understand the nature of the features in the three datasets. Then, minimal dataset pre-processing was carried out on the features. A filter-based feature selection method called ANOVA F-test was used to select promising features that can improve classification performances of the selected learning algorithms. From all the experimental analyses, findings revealed that Random Forest has the overall best performances compared to the two other classifiers. Moreso, all the three algorithms had best average performances in the four metrics in the dataset provided by Tan (2018). The comparative technique used in this study is believed to help provide further insights in phishing detection researches.**

Keywords: Phishing Classification, Phishing Datasets, Cyber Space Attacks, Learning Algorithms

I. INTRODUCTION

As an evidence that phishing-based attacks are growing in unprecedented rate, APWG Fourth Quarter report of 2020 [1] claimed that there is a double increase in phishing attacks. Attackers

amoyelakin@alhikmah.edu.ng

use uniform resource locators or infected file attachments to launch phishing-based attacks. Malicious cyber actors are not relenting in their efforts to launch attacks in the internet space. The financial institutions, webmail, and SaaS site category was the one most frequently victimized by phishing in the fourth quarter of 2020 [1].

Phishing attacks keep using an array of deception techniques to fool users. These include domain names chosen to avoid detection encryption designed to lull victims into a false sense of security. Various machine learning-based approaches have been identified to be promising for identification of cyber threats [2]. One of the key challenges in phishing study is the availability of benchmark datasets that is generally acceptable. This study focuses on investigating the performances of some supervised learning algorithms across three different phishing url datasets. The motive behind this is to provide further insights in phishing-based studies.

As reported in different literature, examples of supervised machine Learning algorithms that can be used for classification or regression tasks include: Random Forest Algorithm, Support Vector Machines, K Nearest Neighbour, ExtraTrees Algorithm, Adaboost, Voting Classifier, XGBoost and so on. It has been argued that the ensemble classifiers are generally more accurate than any of the individual classifiers ([3];[4];[5];[6];[7]). This study changed the direction of comparative studies of machine learning classifiers in phishing detection. The focus of this study is to properly identify phishing evidence from each of the datasets using Random Forest Algorithm, K-Nearest Neighbour and Extra Tree Algorithm. Thereafter, the comparisons of the algorithms are carried out in respect of each of the datasets.

II RELATED STUDIES

Writers in[8]built 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. These results can be said to be low in view of the need to detect every phishing attack in the internet space.

Similarly, the researcher in [9] 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. Authors in [10] carried out a comprehensive review of literature on phishing attack detection. The study focused on the need to provide anti-phishing training 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 solution.

Authors in [11] developed machine leaning based phishing detection models that were validated using a phishing dataset. The focus of the work is to identify how the selected leaning algorithms behave in the only dataset used for the experimentations. Similarly, researchers in [12] 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. Authors in [13]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.

III METHODOLOGY

The methodology used in this work is machine learning-based. The methodology is divided into various stages, which include: data collection, data cleaning, feature selection, and phishing classification using the selected algorithms. The processes are briefly discussed below.

Datasets Collection Process

The first two datasets were obtained from UCI Machine Learning repository while the third one was downloaded from Mendeley data. The datasets were obtained through the links provided herein. For instance, the first dataset in the table 1 is publicly available at https://archive.ics.uci.edu/ml/machine-learning-databases/00379/ as released by [14]. Furthermore, the dataset in table 1 with serial number 2 was obtained from https://archive.ics.uci.edu/ml/machine-learning-databases/00327// as made available by [15]. The third dataset which is a phishing dataset released by [16] is available at http://dx.doi.org/10.17632/h3cgnj8hft.1. The characteristics of the datasets are as shown in table 1. Exploratory Data Analysis was then carried out in order to understand the nature of the features in the three datasets.

About the Datasets

Table 1: Characteristics of the Datasets used

S/N	Dataset Author/Year	No of Input Features	Number of Instances	Missing values?	Data type of Features (Input and Target)
1	Abdelhamid et al. (2014) [14]	9	1353	No	The input features are integer type while the target feature is categorical.
2	Mohammad, McCluskey and Thabtah (2014) [15]	30	11054	No	The input features are integer type while the target feature is categorical
3	Tan (2018) [16]	48	10,000	No	The input features are numeric (integer and floating point) while the target feature is categorical.

Data Preprocessing and Feature Selection Method Used

The data pre-processing steps carried out in this study are meant to make the features in the dataset to be in usable format for the learning algorithms. This is in learning with the argument by [17] on the need to pre-process features to be used for building machine learning models. First

of all, summary statistics was used to ascertain the kind of statistical description contained in each of the datasets. Moreover, the multi-class nature of first dataset was addressed using One-Over-the-Rest technique. Thus, the target variable was turned to binary (phishing and non-phishing). The two other datasets have binary classes by default.

Then, the values in the input features of the datasets were scaled using min-max scalar so as to improve the performances of the proposed learning algorithms. Generally, the formal description of scaling features in a dataset is as follows: Let x be an individual feature value (i.e., a value of the feature in some data point), and min(x) and max(x), respectively, be the minimum and maximum values of this feature over the entire dataset. Min-max scaling technique is used to squeeze (or stretch) all feature values to be within the range of [0, 1].

Mathematically, min max normalisation is: $X_{norm}=(X-X_{min})/X_{max}-X_{min}$ (1)

Thereafter, ANOVA F-score technique was used for selecting most promising features in each of the datasets. The choice of the feature selection method is based on the numerical data as input feature and categorical data as target class. Generally, feature selection is the process of reducing the number of input variables when developing a predictive model [17]. The features with the high scores were taken to be promising for the training of the models.

Experimental Analyses

Working Environment

The working environments in this study are: hardware and software environments. Several runs were carried out during the experimentation and the results obtained are reported in this section.

Hardware

The hardware configuration of the system used for the predictive analysis is as follows: Dual core processor, 4GB RAM, and 500 GB Hard Disk Drive.

Software

The softwares include: Windows 10, Anaconda Python IDE with Python 3.7.2 version and libraries such as Pandas, Sklearn, Numpy and Scipy.

Metrics for Evaluation

The mathematical formulae used for obtaining the values of the performance metrics are:

Accuracy= (TP+TN)/(TP+TN+FP+FN) (2)

Precision=TP/(TP+FP) (3)

Recall=TP/(TP+FN) (4)

F1-score=2×(Precision X Recall)/(Precision + Recall) (5)

IV RESULTS AND DISCUSSION

Table 2: Summary Statistics of the three datasets

S/N	Summary Statistics of the Datasets
1	1 -1 ... 0 0.1 count 1352.000000 1352.000000 ... 1352.000000 1352.000000 mean 0.237426 -0.258136 ... 0.114645 -0.113905 std 0.916493 0.679024 ... 0.318711 0.955121 min -1.000000 -1.000000 ... 0.000000 -1.000000 25% -1.000000 -1.000000 ... 0.000000 -1.000000 50% 1.000000 0.000000 ... 0.000000 -1.000000 75% 1.000000 0.000000 ... 0.000000 1.000000 max 1.000000 1.000000 ... 1.000000 1.000000
2	-1 1 ... -1.15 -1.16 count 11054.000000 11054.000000 ... 11054.000000 11054.000000 mean 0.313914 -0.633345 ... 0.719739 0.113986 std 0.949495 0.765973 ... 0.694276 0.993527 min -1.000000 -1.000000 ... -1.000000 -1.000000 25% -1.000000 -1.000000 ... 1.000000 -1.000000 50% 1.000000 -1.000000 ... 1.000000 1.000000 75% 1.000000 -1.000000 ... 1.000000 1.000000 max 1.000000 1.000000 ... 1.000000 1.000000
3	3 1 ... 1.8 1.9 count 9999.000000 9999.000000 ... 9999.000000 9999.000000 mean 2.445045 0.586759 ... 0.314031 0.499950 std 1.346892 0.751240 ... 0.897862 0.500025 min 1.000000 0.000000 ... -1.000000 0.000000 25% 2.000000 0.000000 ... -1.000000 0.000000 50% 2.000000 1.000000 ... 1.000000 0.000000 75% 3.000000 1.000000 ... 1.000000 1.000000 max 21.000000 14.000000 ... 1.000000 1.000000

The summary statistics provides better insights about the three chosen datasets by exploring them.

No of Selected Features

The filter-based feature selection algorithm chosen in this study is named ANOVA F-test. The features that recorded the highest scores were treated as the promising ones. In each of the datasets, the number of selected variables is as shown in table 3.

Table 3: No of selected features

S/N	Dataset Author and Year	Feature Selection Technique	No of Features Selected
1	Abdelhamid et al (2014) [14]	ANOVA F-score	3
2	Mohammad, McCluskey and Thabtah (2014) [15]	ANOVA F-score	9
3	Tan (2018) [16]	ANOVA F-score	22

Results of Phishing classification using the selected algorithms

The performances of the three machine learning algorithms used in the study were validated using the three datasets described in table 1. For each of the experimentations, the datasets were split in the ratio 80:20 for the training and test sets respectively. Tables 4, 5, and 6 were used to present the performances based on the experimental analyses carried out.

Table 4: Classifier Performances using dataset by [14]

Algorithm	Dataset Used	Accuracy	Precision	Recall	F1-Score
Random Forest Classifier	Abdelhamid et al (2014) [14]	0.91	0.91	0.90	0.90
K-Nearest Neighbour Classifier	Abdelhamid et al (2014) [14]	0.89	0.89	0.89	0.88
ExtraTree Classifier	Abdelhamid et al (2014) [14]	0.90	0.89	0.90	0.89

Table 5: Classifier Performances using dataset by[15]

Algorithm	Dataset Used	Accuracy	Precision	Recall	F1-Score
Random Forest Classifier	Mohammad et al. (2014) [15]	1.00	0.99	0.99	0.99
K-Nearest Neighbour Classifier	Mohammad et al. (2014) [15]	0.92	0.92	0.92	0.92
ExtraTree Classifier	Mohammad et al. (2014) [15]	0.98	0.98	0.98	0.98

Table 6: Classifier Performances using dataset by [16]

Algorithm	Dataset Used	Accuracy	Precision	Recall	F1-Score
Random Forest Classifier	Tan (2018) [16]	0.99	0.97	0.98	0.97
K-Nearest Neighbour Classifier	Tan (2018) [16]	0.97	0.95	0.97	0.96
ExtraTree Classifier	Tan (2018) [16]	0.98	0.96	0.98	0.96

V DISCUSSIONS

The results of the experimental analyses are discussed in this section. Exploratory Data Analysis was first of all carried out in all the three phishing datasets in order to understand the nature of the features in the datasets. This work pre-processed the phishing datasets with a view to making the features available in usable form for the learning algorithms. A filter-based attribute selection technique named ANOVA-F test was used. The chosen feature selection method selected promising features. Afterward, three machine learning based phishing detection models were built. From the experimental analyses, it is evident that Random Forest Classifier has the overall best performances across the four metrics. For instance, from the table 1, Random Forest Algorithm marginally performed better than the two other classifiers. Similarly, from table2, Random Forest algorithm performed better than the two algorithms while Extra Tree Classifier performed better than K-Nearest Neighbour. In table 3, the performances of the three classifiers are promising and Random Forest algorithm did well than the other two marginally under accuracy and f1-score as metrics. It was equally observed that the three algorithms averagely performed better when evaluated with the dataset released by [16] in all the four chosen metrics. Generally, all the machine learning algorithms performed excellently when compared to similar studies.

VI CONCLUSION

This study investigated the performances of Random Forest, K-Nearest Neighbour and Extra Tree algorithms in three publicly available benchmark datasets. The three chosen datasets were pre-processed under the same conditions and were used to train and test the selected supervised learning algorithms. Accuracy, Precision, Recall and F1 score were used as metrics in the performance evaluation. The study reported that Random Forest algorithm has the most promising performances in all the metrics used for the evaluation. Similarly, experimental analyses showed that all the three algorithms have overall performances while using the phishing dataset released by Tan (2018). It is believed that the approach used in this work provided further insights into phishing detection research.

VII ACKNOWLEDGMENT

We would like to thank all the anonymous reviewers who helped improved the original manuscript.

VIII REFERENCES

[1] APWG,"Phishing Activity Trends Report for Quarter four of 2020", APWG Quarterly Report, retrieved from https://docs.apwg.org/reports/apwg_trends_report_q1_2020.pdf

[2] A. Pektaş & T. Acarman, "Botnet detection based on network flow summary and deep learning". *International Journal of Network Management*, 28(6),2018 1–15. https://doi.org/10.1002/nem.2039

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

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

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

[6] I. Breiman, L. ,"Stacked regressions. Machine Learning'" 24(1),1996, 49–64.

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

[8] S. Patil, Y. Shetye, N.Shendage, "Detecting Phishing Websites Using Machine Learning", *International Research Journal of Engineering and Technology (IRJET),2020*,7(2)

[9] V. Shahrivari. "Phishing Detection Using Machine Learning Techniques", arXiv:2009.11116v1 [cs.CR], 2020

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

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

[12] 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),2019, 1–6

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

[15] R. Mohammad, T. L. McCluskey, and Thabtah, FadiAbdeljaber, " 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/

[14] N. Abdelhamid, A Ayesh and F. Thabtah "Phishing detection based Associative Classification data mining. Expert Systems with Applications", 41 (13), 2014, 5948–5959, available at https://archive.ics.uci.edu/ml/machine-learning-databases/00379/

[16] C. L. Tan, Choon Lin, " Phishing Dataset for Machine Learning: Feature Evaluation, Mendeley Data, V1, 2018, doi: 10.17632/h3cgnj8hft.1, the dataset is available at http://dx.doi.org/10.17632/h3cgnj8hft.1

[17] M. A. Hall, "Correlation-based Feature Selection for Machine Learning", *a PhD Thesis at University of Waikato, 1999*