A COMPARATIVE ANALYSIS OF DIFFERENT FEATURE SET ON THE PERFORMANCE OF DIFFERENT ALGORITHMS IN PHISHING WEBSITE DETECTION

Hajara Musa[1], Bala Modi[2], Ismail Abdulkarim Adamu[3], Ali Ahmad Aminu[4], Hussaini Adamu[5], Yahaya Ajiya[6].

[1, 2,4,5,6] Department of Mathematics, Faculty of Science, Gombe State University, Gombe, Nigeria

[3]Department of Computer Science, School of Science Technology, Gombe State Polytechnic, Bajoga

ABSTRACT

Reducing the risk pose by phishers and other cybercriminals in the cyber space requires a robust and automatic means of detecting phishing websites, since the culprits are constantly coming up with new techniques of achieving their goals almost on daily basis. Phishers are constantly evolving the methods they used for luring user to revealing their sensitive information. Many methods have been proposed in past for phishing detection. But the quest for better solution is still on. This research covers the development of phishing website model based on different algorithms with different set of features in order to investigate the most significant features in the dataset.

KEYWORD:

Machine learning, Feature selection, Phishing, XGBoost, Random Forest (RF) and Probabilistic Neural Network (PNN).

1. INTRODUCTION

According [1] the word "phishing" is coined from the word "fishing". Phishing is a criminal activity that takes users' own information using deceptive emails, or fake websites addresses. Online internet users can be simply be deceived into giving their private information because phishing websites are highly similar to real ones.

Phishing is a cyber-crime which involves the fraudulent act of illegally capturing private information like credit card details, usernames, password, account information by pretending to be authentic and esteemed in instant messaging, email and various other communication channels. The traditional approaches used by majority of the email filters for identifying these emails are static which make it weak to deal with latest developing patterns of phishing since, the defrauders are dynamic in actions and keep on modifying their activities to dodge any kind of detection[2].

Phishers operate by sending fake emails to their victims pretending to be from legitimate and well known organizations such as banks, university, communication network etc. where they will require updating some personal information including their passwords and usernames to avoid

losing access right to some of the services provided by that organization. Phishers use this avenue to obtained users sensitive information which they in turn use it to access their important accounts resulting in identity theft and financial loss [3].

Many approaches have been proposed in an attempt to curb the problems caused by phishers [4]. However, due to the dynamic nature of attackers and the challenging nature of the problem, it still lacks a complete solution. Recently, machine learning approaches have been found to be very successful in automated detection of phishing web sites. This research work capitalized on this by using XGboost (an optimized implementation of gradient boosted decision tree algorithm) to improve the performance that a predictive model can achieve in the detection of a phishing website from a legitimate website. The paper is organized as follows: Section 2 presents some related work while Section 3 describes the methodology of our approach. Section 4 presents an evaluation criteria. Experiment and discussion of result is presented in Section 5 and finally Section 6 concludes the paper and suggests future work.

2. RELATED WORK

In a research conducted by [3], they investigated the problem of website phishing using a developed AC method called Multi-label Classifier based Associative Classification (MCAC) to seek its applicability to the phishing problem. They also want to identify features that distinguish phishing websites from legitimate ones. Experimental results using real data collected from different sources show that AC particularly MCAC detects phishing websites with higher accuracy than other intelligent algorithms. The problem of the approach is that, many algorithms suffer from defects to varying degrees. It is obviously imperative to achieve correct prediction but also equally or perhaps more important to avoid false and potentially misleading ones. Researchers in [5] proposed heuristic-based phishing detection technique that employs URL-based features. The system first extracts the features which clearly differentiate that whether website are phished or legitimate. The experiment shows that SVM has accuracy of 96% and very low false-positive rate. The proposed model can reduce damage caused by phishing attacks, because it can detect new and temporary phishing sites. Heuristic evaluation does not allow a way to assess the quality of redesigns.

In a recent work conducted by [6] they proposed a hybrid model to classify phishing emails using machine learning algorithms with the aspiration of developing an ensemble model for email classification with improved accuracy. They have used the content of emails and extracted 47 features from it. Going through experiments, it is observed and inferred that Bayesian net classification model when ensemble with CART gives highest test accuracy of 99.32%. The approach creates over-complex trees that do not generalize the data well is called over fitting.

Scholars in [7] compare different features assessment techniques in the website phishing context in order to determine the minimal set of features for detecting phishing activities. Experimental results on real phishing datasets consisting of 30 features has been conducted using three known features selection methods. Their approach can be hard to find a usable formal representation and it deals badly with quantitative measurements. The emails have been classified as phish using the prediction of Ensemble Classifier of the five ML Algorithms in [2] experiment shows that the comparison of the accuracy of algorithms for Different Feature Groups based on the decisive values of the features demonstrated that best accuracy is obtained for Random Forest by 96.07%. Random forests have been observed to over fit for some datasets with noisy classification tasks. The evaluation of model size is slow because it could easily end up with a forest that takes hundreds of megabytes of memory.

In a work of [8] they presented a novel approach for detecting phishing websites based on probabilistic neural networks (PNNs). They also investigate the integration of PNN with K-means clustering to significantly reduce complexity without jeopardizing the detection accuracy. The experimental results show that 96.79% accuracy is achieved with low false errors. Their approach requires large memory spaces to store and the execution of network of this approach is slow.

Phishing is a continuous problem. Thus, there is a need to constantly improve the network structure in order to cope with these changes [9] the quest for a better solution is still on. In recent time, machine learning techniques have been found to be very successful in phishing website detection [10][11][12] .This proposes XGBOOST algorithm to improve the performance that a predictive model can achieve in the task of phishing website detection. Advantages of XGBOOST have made it an excellent tool of choice for many researchers in data science and machine learning. In light of the above, XGBOOST has been recently employed in many machine learning task with great success [13][14][15].

3. METHODOLOGY

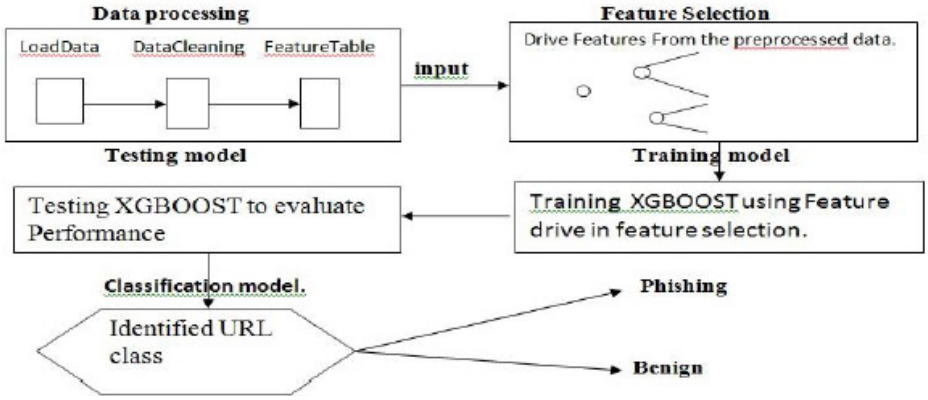

Figure1. illustrate the framework of the proposed model

In order to test the effect of the dataset size, feature selection is important because dataset may contain irrelevant noisy and redundancy feature in which if they are included (incorporated), it will surely affect the model negatively. Feature selection is one of the data mining techniques used in data pre-processing stage.

Firstly, the relevant datasets are collected and pre-processed before being fed into the proposed model for training and testing. Finally, the model is evaluated based on standard evaluation metrics and the model classified either the website is benign or phishing.

To investigate and compare the performance of the proposed model, experiment was conducted using a benchmark phishing website dataset created by [16] (Mohammad et al, 2014).

4. EVALUATION CRITERIA

To evaluate and compare the performance of different features categories we have to measure the accuracy (ACC) , precision (Prec), recall (Rec), mathew correlation coefficient (MCC), false positive rate (FPR), false negative rate (FNR) and f-score. ACC measures the ratio of websites which are correctly predicted. Prec measures the fraction of websites correctly predicted as phishing. Rec metric measures the fraction of phishing websites identifield by the model. MCC

measures the correlation coefficient between the predicted and actual class. FPR measures the % of not faulty websites labeled as fault prone by the model. FNR measures the % of faulty websites labelled as not faulty by the model. F-score measures the weighted hormonic mean of precision and recall. All metrics employed are functions of the confusion matrix as can be seen in the mathematical formulatons. The confusion matrix shown in table1 is a table use to describe the performance of a classification model on a set of test data for which the true values are given.

Table1. Confusion matrix

	Predicted positive class	Predicted negative class
Actual positive class	TP	FN
Actual negative class	FP	TN

The abbreviations TP, FN, FP and TN in table1 are explained below respectively.TP (True Positive) is a case where a model correctly predicts a website as phishing, TN (True Negative) is a case where a website is wrongly classified as benign. FP (False Positive) is a case where a website is wrongly classified as phishing and lastly FN (False negative) is when the model wrongly classified a website as benign while it is actually phishing. The mathematical equations of the performance metrics are given below respectively.

$$ACC = \frac{TP + TN}{(TP+TN+FP+FN)} \quad (1)$$

$$Prec = \frac{TP}{(TP+FP)} \quad (2)$$

$$Rec = \frac{TP}{(TP+FN)} \quad (3)$$

$$F - score = \frac{2*(Prec*Rec)}{(Rec+Prec)} \quad (4)$$

$$MCC = \frac{(TP*TN)-(FP*TN)}{Sqrt(TP+FP)(TP+FN)(TN+FP)(TN+FN)} \quad (5)$$

5. EXPERIMENT RESULT

The experiment was carried out using different set of features in order to investigate the most significant features in the dataset. Table2 contained the result of six (6) subset of feature.

Table2. Results of RF, PNN, and XGBOOST using feature selection

Feature category	Precision	Recall	F.score	MCC	Accuracy
RF,Address bar based features (1–12)	0.8986	0.9411	0.9194	0.8176	0.9096
PNN,Address bar based features (1–12)	0.8717	0.9163	0.8935	0.7531	0.8783
Xgboost,Address	0.9106	0.9283	0.9096	0.8193	0.9111

bar based features (1–12)					
RF,Abnormal based features (13–18)	0.8271	0.9576	0.8876	0.7388	0.867
PNN,Abnormal based features (13–18)	0.8191	0.8757	0.8465	0.6405	0.8232
Xgboost,Abnormal based features (13–18)	0.8883	0.9585	0.8702	0.7519	0.8752
RF,HTML & JavaScript based features (19–23)	0.5566	0.9851	0.7113	0.9704	0.5616
PNN,HTML & JavaScript based features (19–23)	0.5646	0.9935	0.72	0.11	0.5699
Xgboost, HTML & JavaScript based features (19–23)	0.6583	0.9859	0.4100	0.1106	0.5746
RF,Domain-based features (24–30)	0.7581	0.7635	0.7608	0.4683	0.7368
PNN,Domain-based features (24–30)	0.6936	0.8863	0.7782	0.4325	0.7187
Xgboost,Domain-based features (24–30)	0.7392	0.8032	0.7351	0.4725	0.7416
RF,Feature selection (6–8, 13–16, 26, 28)	0.9405	0.9565	0.9484	0.8849	0.9450
PNN,Feature selection (6–8, 13–16, 26, 28)	0.9245	0.935	0.9297	0.8404	0.9213
Xgboost,Feature selection (6–8, 13–16, 26, 28)	0.9416	0.9515	0.9412	0.8825	0.9421
RF,Full dataset (1–30)	0.9433	0.9796	0.9611	0.9128	0.9566
PNN,Full dataset (1–30)	0.9576	0.9724	0.9649	0.9203	0.9607
XGBOOSTFull (1–30)	0.9730	0.9801	0.9724	0.9449	0.9729

Table 2 shows the result obtained from the experiment using six feature categories based on the accuracy in which HTML and javascript based features has 0.5746, domain based feature has 0.7416, abnormal based feature has 0.8752, address bar based feature has 0.9111, feature selection has 0.9421 and full dataset has 0.9729. Therefore, the result shows that the full dataset is better campared with others feature categories.

This result can be represented in a graphical form for easy analysis. Figure2 shows the representation of the proposed Model chart using different feature categories of the dataset.

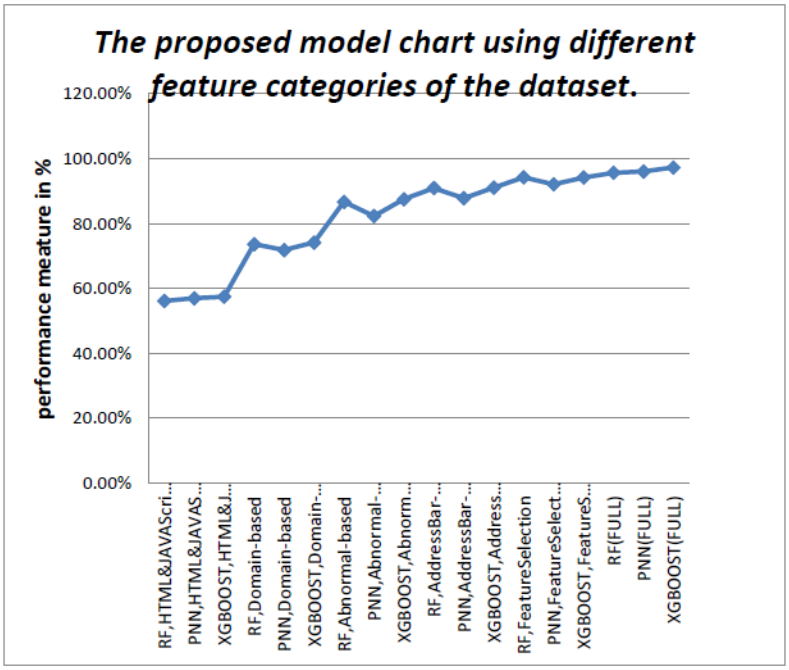

Figure2. The proposed Model chart using different feature categories of the dataset

Figure 2 demostrates the result obtained after applying feature selection method utilizing six(6) categories of subset features which are: address bar based feature, abnormal based feature, domain based feature, feature selection, HTML and javascript based features, full dataset. The results shows that the collection of address bar based attained >91% accuracy while that of PNN is >87% accuracy. Using feature selection with nine (9) subsets features, the performance of XGBOOST achieved >94% accuracy while that of PNN returned 92% accuracy. But incase of HTML and javascript based feature both XGBOOST and PNN has very poor performance results with 57.46% and 56.99% respectively. This demostrated that using full dataset is better because it generate and returned high accuracy performance which indicate that the combination of all features is important.

6. CONCLUSION

Conclusively, this work has shown that XGBOOST can be adapted to obtain a very impressive result in detecting phishing. The performance of XGBOOST has been compared with that of well-known techniques Random forest and probabilistic neural network. The evaluation criteria are used in measuring the performance of phishing detection. Benchmark phishing website dataset were considered in the experiment. The result of the experiments showed that XGBOOST is better in most of the problems than the other methods in terms of the F.score, MCC, and Accuracy. Therefore, the xgboost method represents a very competitive technique for phishing detection. XGBOOST has a better regularization ability which helps to reduce overfitting, high speed and performance owing to the parallel nature in which trees are built, flexibility due to it costume optimization objectives and evaluation criteria, and inbuilt routines for handling missing values which makes it good classification algorithm. In view of that, for future work, the application of XGBOOST will be applied on more complex classification problems.

REFERENCES

[1] C. Pham, L.A.T, Nguyen, N.H. Tran, E.N. Huh and C.S. Hong." Phishing-Aware: A Neuro-Fuzzy Approach for Anti-Phishing on Fog Networks". IEEE Transactions on Network and Service Management. Vol. (15), 3, pp. 1076-1089, April 2018

[2] D. P. Yadav., P. Paliwal, D. Kuma, and R. Tripathi, "A Novel Ensemble Based Identification of Phishing E-Mails" In Proceedings of the 9th International Conference on Machine Learning and Computing Singapore. .pp. 447-45, Feb 2017

[3] N. Abdelhamid, A. Ayesh, and F. Thabtah," Phishing detection based Associative Classification data mining". Journal of Expert Systems with Applications, ELSERVIER Vol 41(13), pp.5948–5959. doi:10.1016/j.eswa.2014.03.019.

[4] S. Abu-Nimeh, D. Nappa,X. Wang, and S. Nair. "A comparison of Machine learning techniques for phishing detection" In Proceedings of the anti-phishing working group 2nd annual eCrime researchers' summit. ACM, pp. 60-69, 2007

[5] J. Solanki, and R.G. Vaishnay. "Website Phishing Detection using Heuristic Based Approach". International Research Journal of Engineering and Technology (IRJET), Vol. 3, pp.2044–2048, May 2016

[6] N. Vaishnaw, S.R. Tandan, M.T Scholar and C.G. Bilaspur. "Development of Anti-Phishing Model for Classification of Phishing E-mail" International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4(6), pp. 39-45, June 2015. doi:10.17148/IJARCCE.2015.4610.

[7] F. Thabtah, and N. Abdelhamid. "Deriving Correlated Sets of Website Features for Phishing Detection: A Computational Intelligence Approach." Journal of Information & Knowledge Management. Vol. 15(4), pp. 1–17. Doi: 10.1142/S0219649216500428.

[8] E.M. El-Alfy. "Detection of Phishing Websites Based on Probabilistic Neural Networks and K-Medoids Clustering." The Computer Journal.Vol. 60(12), pp. 1745-1759.Dec 2017.

[9] R.A. Mohammad. F. Thabtan,and L. Mccluskey. (2014). "Predicting phishing websites Based on self-structuring neural network." Neural Computing and Application. Vol. 25(2), pp. 443-458, Dec 2013 doi: 10.1007/s00521-013-1490

[10]M. Kaytan and D. Hanbay. "Effective Classification of Phishing Web Pages Based on New Rules by Using Extreme Learning Machines". Anatolian Journal of Computer Sciences .Vol. 2(1), pp. 15-36, 2017.

[11] K.B. Kazemian and S. Ahmed. (2015)." Comparisons of machine learning techniques for Detecting malicious webpages" Expert Systems with Applications, Vol. 42(3), pp. 1166-1177, Feb 2015

[12] A.K. Jain and B.B. Gupta. "Comparative analysis of features based machine Learning approaches for phishing detection" In 3rd International Conference on Computing for Sustainable Global Development (INDIACom), IEEE pp. 2125-2130, Oct 2016.

[13] T. Zimmermann, T. Djürken,A. Mayer,M. Janke, M. Boissier,C. Schwarz and M. Uflacker, "Detecting Fraudulent Advertisements on a Large E-Commerce Platform". In EDBT/ICDT Workshops.2016.

[14] X. Wei, F. Jiang, F. Wei,J. Zhang, W. Liao and S. Cheng."An Ensemble Model for Diabetes Diagnosis in Large-scale and Imbalanced Dataset". In Proceedings of the Computing Frontiers Conference, ACM, pp. 71-78. May 2017.

[15] L. Zhang and C. Zhan. “Machine Learning in Rock Facies Classification: An Application of XGBOOST”. In International Geophysical Conference, pp. 17-20 April 2017.

[16] R.M. Mohammad, F. Thabtah and L. McCluskey.” Intelligent rule-based phishing websites classification”. IET Information Security. Vol. (8), 3, pp. 153–160, May 2014.