Phishing Detection Using Machine Learning Algorithm.

Jibrilla Tanimu
Department of Computing
University of Portsmouth
jibrilla.tanimu@port.ac.uk

Stavros Shiaeles
Department of Computing
University of Portsmouth
sshiaeles@ieee.org

2022 IEEE International Conference on Cyber Security and Resilience (CSR) | 978-1-6654-9952-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/CSR54599.2022.9850316

***Abstract* - The rapid increase of phishing attacks has led individuals and organizations losing billions of dollars as well as worried about the confidentiality and privacy of their data. This tremendous annual increase of phishing attacks shows that the current detection methods available are not sufficient, therefore more effective phishing detection methods should be developed. This paper proposed a novel phishing detection model using machine learning, to improve efficacy and accuracy in phishing detection. This paper explores the current state-of-the-art in phishing detection along with their drawbacks and proposes a new novel method based on image visualisation of website code and features extraction from malicious URLs which is under development.**

Index Terms* - *Machine learning, Phishing detection, Images, Binary visualisations, Spam

I. INTRODUCTION

Companies are losing 100 billion dollars per year because of phishing and even worse, phishing attacks are increasing every year by 200%. This leads to a conclusion that the current solutions available are not adequate, and new methods should be developed to protect companies and end-users. Many financial activities are computerized and less cash is available which led to the new trend of phishing and other cybercrimes, defrauding internet users is the new trend, to collect their financial credentials. In recent times many criminal organizations have shifted from exploiting system vulnerabilities in information systems to human's inability to differentiate between legitimate and fake online resources such as email and websites. This makes it significantly important to provide a solution to mitigate the problems.

The need for an efficient solution has made phishing detection a favourable research area in recent times, with the unfolding of the visual similarities approach, blacklisting and whitelisting approach, website content and URL features. The visual similarities utilise extraction of features from a website and re-use them for phishing website identification; this method is ineffective as any misrepresentation of a web page's content affects the visual content leading to misclassification. The blacklisting and whitelisting are also ineffective in classifying phishing and non-phishing website; however, the system cannot understand a newly developed website that is not registered, and this could also contribute to misclassification. The current innovative approach to classifying phishing activity is the utilisation of URL and web content features with a machine learning approach to improve detection accuracy and performance.

Alabdan [1] suggested that frequent system updates and regular training of staff on new phishing activities will massively reduce the number of phishing attacks in an organization.

In this paper, we aim to provide a solution to the limitations aforementioned in the related works by proposing an extensive phishing detection using a machine learning approach. Using image classification with some selected features, such as domain base features, address bar-based features, abnormal-based features and HTML base features combined with phishing website URLs to create an image used in our proposed model to perform prediction with different classification algorithms. To archive high efficiency and efficacy.

The paper is organised as follows: Section 2 provides the related work done in the research area and what needs to be done (drawback). While in Section 3, we present the proposed method, the feature that will be used along with the image classification, while Section 4 presents the stage that has been implemented and finally, Section 5, concludes the paper and presents the future work.

II. RELATED WORK

Different methods and approaches have been explored to understand and provide a solution to phishing attacks. In this respect, many papers are available that examine strategies adopted by phishers or attackers, but here we will be focused on the most relevant papers that provide the best accuracy. Barlow et al. [2] present a comprehensive novel approach for detecting phishing attacks using binary visualisation and machine learning. The authors highlight the need for fast access time so that the proposed method to be functional in a real-time environment and also have a high detection rate. This was achieved by combining phishing threats with binary

978-1-6654-9952-1/22/$31.00 ©2022 IEEE

visualisation and machine learning. Adebowale et al. [3] present an article on different types and methods of phishing attacks as well as new ways to mitigate them. The article claims to have achieved 95.83% accuracy in detecting phishing attacks.

Luga et al. [4] Analyse data to understand the percentage of users that fall victim to phishing attacks by analysing data and finding that 65.63% of users will undoubtedly fall victim to phishing scammers. Some of the data used were Gender, Computer manufacture date, Operating system, etc. in addition to individual page course analysis with correlated detection scores.

Almseidin et al. [5] implement a prototype for phishing detection using machine learning and testing it with 500 fake and 500 legitimate web pages. The paper concludes that Random Forest is the best classifier to be used as it was able to detect a phishing attack in 2.44 seconds and with 98.11 accuracy. The article filter the assigned data to increase performance and reduce computation time. Some of the classifiers used in this research work are Net, Naive Bayes J48, Logistic, Random Forest, Bagging and multiple perceptions.

Gualberto et al. [6] discuss different email phishing attacks and focused on the natural language processing (NLP) technique in conjunction with the machine learning approach. The authors achieved a 99.95 accuracy using XGBoost Algorithm. Furthermore, the authors dealt and compared with different classifiers such as Multilayer perception (MLP), Support Vector Machines (SVM), Logic regression for Classification (LRC), K-Nearest Neighbour (KNN), Random Forest (RF), Decision Tree (DT), Extreme Gradient Boosting (XGBoost) and Multilayer perception (MLP). The article summarised phishing detection techniques based on natural language technique and machine learning in a data-driven approach presented to be of great effectiveness and higher accuracy than those depending on filtering rules and also augured that their proposed solution is very effective and good enough to identify and mitigate phishing attacks.

According to [7] a novel model for the detection of phishing websites similar to Kumar & Gupter [8] is proposed. Both papers are using features from websites in order to detect phishing and they compare their result with CANTINA and CANTINA+ to show the effectiveness of their proposed model. The models identify phishing sites on features extraction from URL content and other third-party resources such as machine learning algorithms. According to the authors, the proposed method achieved a 99.55% detection rate and a very low false-positive rate of 0.45% utilising the Random Forest Algorithm. The research uses a mode that uses phishing sites, which is a replica of the legitimate site by placing some of the content with an image which they believe a lot of research articles neglect to put into consideration.

Also, Basit et al [9] present a prototype for detecting phishing attacks on a website using the Machine learning algorithm by examining three major machine learning classifiers which are: K- Nearest Neighbour (K-NN), Artificial Neuron Network (ANN) and Decision Tree (C4.5) to cast with Random Forest Classifiers (RFC). The paper suggests RFC has the highest detection accuracy reaching 97.33% when compared to other classifiers. For the experiment, 4898 legitimate and 6157 phased websites were used and they also conclude that putting more variables into the process will improve the detection accuracy.

Table 1:Overview of the prior research

Record of survey papers based on classifications					
Reference	***Solutions***	***Approach***	***Limitations***	***Accuracy***	***Future enhancement***
[2]	Phishing and non- non phishing website classification.	Uses the combination of neural network with binary visualization	Dataset use to conduct the experiment was 4,000 which limit the predication and affect the efficacy of the model	95.89%	Adding more dataset for both training and testing and also trying different model in making predictions
[10]	Phishing URL Detection	Introduce TOROEDO to map out existing email phishing problems	IT literate were used to carry out research	85.17% Phishing Detection	Simulation environment need to be available for user testing
[11]	Phishing detection/ offensive defence	User training to increase awareness on phishing	There were no matrices for end user evaluation for each website	Not stated. Increase awareness level of users. Though it acknowledges machine learning as the most	It could also be extension in enhancing anti-phishing training

				prominent approach	
[12]	Automatic data collecting and security analysis	Role in detecting phishing activities using random forest	Some of the processes were not clearly demonstrated till the final stage	95.2% precision and 91.6 recall in phishing detection.	Future research of this kind needs to enhance early detection of phishing activities
[13]	Phishing detection and mitigation in emails/website	Blacklist approach, cloud-threat inspection approach.	A lot phishing detection method such as rule base method, decision tree, associative classification, SVM, NN were listed but none have been demonstrated in the research.	Not applicable It is a literature publication	Incorporate other phishing detection technique such as SVN, K-nearest neighboured in solving
[14]	Cyber security awareness and training	Using place management approach to investigate in cybercrime	Logical analysis revealed very few differences among control variable that were included in the overall research, which increase the possibility of the research not capturing some other variables	Not applicable Awareness and training of individual in phishing detection.	The research generally focusses on training, so the need to continuous train consumers of the internet should be improve and maintain, for feature security implementation in that aspect.
[15]	The architecture and working style of ANFIS and DNN	Deep Neuro fuzzy classifier approach; which were divided in to different segment; MF, fuzzification part, deep learning part.	The result generated by DNN has huge setback as it was not transparent and not easy to understand.	Not applicable clustering using Neuro fuzzy model	To expand the research to see how the issues of not simply to understand is been mitigated.

In summary, researchers extensively involved with the subject tried to mitigate phishing attacks by proposing a novel technique that has a regular detection rate, not much better than the already existing result, also they fail to answer the research question in some cases the result comes with medium accuracy but high false positive.

Some research adapts feature extractions such as visual similarities between phishing and legitimate website in their studies, entities such as text content, Cascading style sheet (CSS), HTML tags, text formatting and so forth, were used to classify between legitimate and phishing websites, most of the time the solution was implemented without considering the importance of classification accuracy or recognition speed. These kinds of approaches can only classify based on the given tag or text formatting and so forth, this kind of classification may lead to ambiguity as the entity that is not defined in the model will not be classified, which will affect the efficacy and efficiency of the classification.

Similarly, some literature shows how extensively the researchers attempt to invest in whitelisting and blacklisting of the domain, to enable them to classify addresses from phishing or legitimate sender, this kind of study is very vulnerable to misclassification and increase the rate of false alarm, in case of new un-registered address, whitelisting and blacklisting cannot classify new unregistered phishing or non-phishing website.

In this research, we aim to address the above-mentioned issues by combining features and HTML code of 80,000 phishing websites to create unique images and train a machine learning model so it could identify malicious websites with high accuracy in real-time. The feature of the proposed solution is provided in the Section below.

III. Proposed Method

This Section provides our proposed model that is currently under development as shown in Fig 1. The first component is a crawler. This component is crawling continuously Phish Tank and storing all the actual Phishing websites in the database. These websites are then parsed with another program where features are collected and constructed into a unique image that is feeding the machine learning in order to provide a decision. Table 2 shows some of the features we found based on our extensive literature review. In Fig. 2 a quick overview of the current phishing URLs is provided and some features can be exported like IPAddress, Second level domain etc.

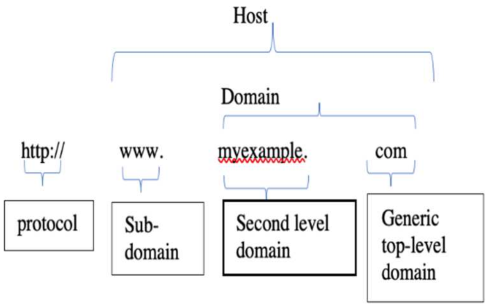


Fig. 2: URL features

During the machine learning progress in order to make the best prediction, we will test different classification algorithms to understand their performance and improve their accuracy by optimising the image creation.

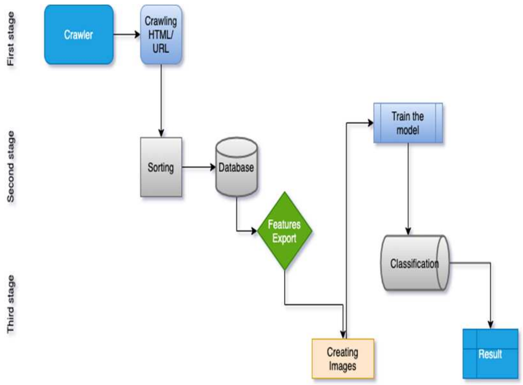


Fig. 1: Phishing Detection Proposed Model

Another important part of the procedure is feature elimination, which would enable us to remove the less significant features that exist in the dataset and improve visualisation as well as picture creation speed. This would also help machine learning to accelerate detection time without compromising detection accuracy. This will be discussed further in the Implementation Section below.

Currently, the following machine learning classifiers are considered based on literature review which is: Decision tree, Support Vector Machine, Naïve base, 2D and 3D Neural Networks, TensorFlow and some newly proposed algorithms such as artificial general intelligence (AGI). Our goal is to improve the detection accuracy and time taken for the classification and compare them with the most promising methods proposed in the literature.

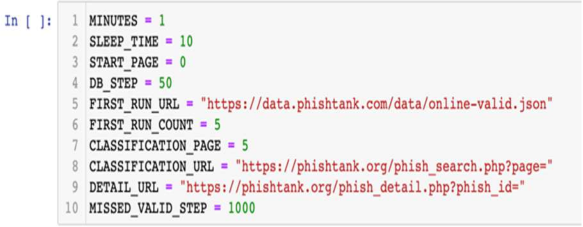
Table 2:Some of the feature of phishing and non-phishing website in the dataset

S/N	Features	Possible values
1	LinksInTags	1,-1,0
2	AbnormalURL	1, -1
3	AgeOf Domain	1,-1
4	Port	1,-1
5	RightClickDisabled	1,-1
6	PopUpWindows	1,-1
7	EmbeddedBrandName	1,-1
8	SubdomainLevel	1,-1
9	Redirect page	1,-1
10	IpAddress	-1,1
11	PctExtResourceUrls	1,-1
12	InsecureForms	-1,1
13	double slash redirecting	-1,1
14	FrequentDomainNameMismatch	1,-1
15	UrlLengthRT	1,-1
16	ExtMetaScriptLinkRT	1,-1
17	Using pop-up window	1,-1
18	DoubleSlashInPath	-1,1
19	MissingTitle	1,-1
20	Page Rank	1,-1
21	SSLfinal State	1,0,-1
22	FakeLinkInStatusBar	1,1-
23	RandomString	-1,1
24	HostnameLength	1,-1
25	QueryLength	1,-1
26	NoHttps	1,-1
27	Links pointing to page	-1,0,1
28	NumHash	1,-1
29	IframeOrFrame	-1,1
30	InsecureForms	-1,1

IV. Implementation

The first step in our proposed method is the crawler which is implemented and is used to continuously retrieve phishing data from Phish Tank and add them into the MySQL database (Fig. 4) for further data manipulation. The crawler can handle a large amount of data, using python programming language running on an Ubuntu 20.04 LTS, 2 cores, 8 GB or Ram Virtual

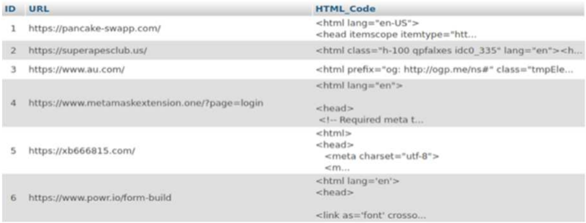
Machine. Python was selected because it makes development faster and easier by utilising packages such as MySQL connector for the database connection, Pandas and other libraries which simplify our work. Libraries such as Scikit-learn, NumPy and Pandas will be utilized as well for the Machine learning implementation.



```
In [ ]: MINUTES = 1
        SLEEP_TIME = 10
        START_PAGE = 0
        DB_STEP = 50
        FIRST_RUN_URL = "https://data.phishtank.com/data/online-valid.json"
        FIRST_RUN_COUNT = 5
        CLASSIFICATION_PAGE = 5
        CLASSIFICATION_URL = "https://phishtank.org/phish_search.php?page="
        DETAIL_URL = "https://phishtank.org/phish_detail.php?phish_id="
        MISSED_VALID_STEP = 1000
```

Fig. 1: Crawler code, collecting Phishing data from phish Tank

For the last months, the crawler was running continuously, collecting both legitimate and phishing websites. The current amount of data reaches 30000 malicious and benign URLs that contain the sub-entities shown in Fig. 2. These data will be used later for the training and testing of our proposed model.



ID	URL	HTML_Code
1	https://pancake-swapp.com/	<html lang="en-US"> <head itemscope itemtype="htt...
2	https://superapesclub.us/	<html class="h-100 qpfalxes idc0_335" lang="en"><h...
3	https://www.au.com/	<html prefix="og: http://ogp.me/ns#" class="tmpEle...
4	https://www.metamaskextension.one/?page=login	<html lang="en"> <head> <!-- Required meta t...
5	https://xb666815.com/	<html> <head> <meta charset="utf-8"> <m...
6	https://www.powr.io/form-build	<html lang='en'> <head> <link as='font' crosso...

Fig. 2: Data from MySQL database showing data collected

The feature Elimination method is the next step in our proposed solution. Currently, this step is not implemented as we are in the state of looking at various methods available, the right sets of features would be determined by the features elimination prosses to select the most significant features. Below are the most prominent methods that have been investigated and will be tested.

Recursive feature elimination

Works by recursively removing attributes and building a model on the remaining attribute, which makes it suitable to remove the redundant phishing attribute in the proposed model.

We represent RFE as:

$$rfe = rfe.fit(X\ train\ Y\ train)$$

The principal component analysis

(PCA) method utilises linear algebra to transform data into a compressed form also known as the *data reduction technique.* Which makes it an essential approach to reduce the less relevant features and save time

We can use the covariance of X and Y using the following:

$$cov\ (X,Y) = \frac{1}{n-1}\sum\nolimits_{i=1}^{n}(Xi - x)(Yi - y)$$

The feature importance method

is a bagged decision tree such as a random forest and extra tree used to estimate the importance of the features and eliminate the less important feature. The importance of each feature in the decision tree is calculated as:

$$fi_i = \frac{\sum j: node\ j\ splits\ on\ feature\ n^i\ j}{\sum_{k\in all\ nodes} ni_k}$$

fi sub (i) = the importance of feature i

ni sub (j) = the importance of node j

Univariate selection method

This approach statistical test would be utilised to select those features that have the strongest features with the output variables, that have a significant impact factor on the feature in the proposed model. The univariate score statistic is:

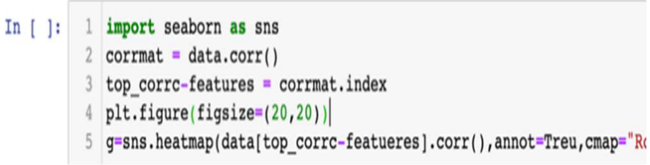
$$S_j = \sum\nolimits_{i=1}^{n} \delta_i\left(x_{ij} - S_{ij}^{(1)}/S_{ij}^{(0)}\right)$$

The heatmap

The approach also enables the visualising feature provided in the model by finding the correlation between all the values, a feature 1 also known as dependant variable y1 takes all the values in feature 2 also known as dependant y2 which takes:

$$\rho x, y = corr(X,Y) = \frac{cov(X,Y)}{\delta x\delta y} = \frac{E[(X - \mu x)(Y - \mu y)]}{\delta x\delta y}$$

As shown in figure 5, is a sample approach to implementing heatmap using python.



```
In [ ]: import seaborn as sns
        corrmat = data.corr()
        top_corrc-features = corrmat.index
        plt.figure(figsize=(20,20))
        g=sns.heatmap(data[top_corrc-featueres].corr(),annot=Treu,cmap="R
```

Fig. 5: Plotting correlation with Heatmap

We aim to test all the aforementioned feature elimination methods in order to narrow down the number of features and select the most appropriate method for the proposed algorithms. This way our image creation process will receive only important data thus we will reduce the image creation time and overfitting.

The final component of the proposed solution is image creation. Images are created based on the features identified above and the intention is to find commonalities between the various website image samples. These images will be used to feed the final machine learning component where we will train and test our neural network utilising the TensorFlow and other machine learning algorithms dedicated to image recognition as identified above. We will compare their output results in terms of accuracy and we will then conclude on which machine learning algorithm we will be focusing our further research in order to achieve higher accuracy.

V. Conclusion and Future Work

Phishing is an important threat in the corporate environment causing many financial losses. Even if many solutions have been proposed and utilised by well-known cyber security companies, the rapid increase of successful phishing attacks is an indicator that is insufficient to tackle the problem. The current phishing detection and mitigation method proposed in this paper will enhance previous work by providing better results and accuracy. The proposed method is still under development and a huge dataset of phishing and legitimate websites is been collected. The next step is to utilise the feature elimination techniques identified in Section IV, which include Recursive feature elimination, Principal component analysis, feature importance and Univariant selection method and proceed with creating images and training our machine learning models to measure the efficiency and efficacy of our new model, compared with previous studies.

Acknowledgement

This project has received funding from Petroleum Technology Development Funds (PTDF) Nigeria. The work reflects the author's view not the organizational view regarding their findings.

References

[1] R. Alabdan, “future internet Phishing Attacks Survey: Types, Vectors, and Technical Approaches”, doi: 10.3390/fi12100168.

[2] L. Barlow, G. Bendiab, S. Shiaeles, and N. Savage, “A Novel Approach to Detect Phishing Attacks using Binary Visualisation and Machine Learning,” in Proceedings - 2020 IEEE World Congress on Services, SERVICES 2020, Oct. 2020, pp. 177–182. doi: 10.1109/SERVICES48979.2020.00046.

[3] M. A. Adebowale, K. T. Lwin, E. Sánchez, and M. A. Hossain, “Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text,” Expert Systems with Applications, vol. 115. Elsevier Ltd, pp. 300–313, Jan. 01, 2019. doi: 10.1016/j.eswa.2018.07.067.

[4] C. Iuga, J. R. C. Nurse, and A. Erola, “Baiting the hook: factors impacting susceptibility to phishing attacks,” Human-centric Computing and Information Sciences, vol. 6, no. 1, Dec. 2016, doi: 10.1186/s13673-016-0065-2.

[5] M. Almseidin, A. M. Abu Zuraiq, M. Al-kasassbeh, and N. Alnidami, “Phishing detection based on machine learning and feature selection methods,” International Journal of Interactive Mobile Technologies, vol. 13, no. 12, pp. 71–183, 2019, doi: 10.3991/ijim.v13i12.11411.

[6] E. S. Gualberto, R. T. de Sousa, T. P. B. de Vieira, J. P. C. L. da Costa, and C. G. Duque, “From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection,” IEEE Access, vol. 8, pp. 76368–76385, 2020, doi: 10.1109/ACCESS.2020.2989126.

[7] R. S. Rao, • Alwyn, and R. Pais, “Detection of phishing websites using an efficient feature-based machine learning framework,” Neural Computing and Applications, vol. 31, doi: 10.1007/s00521-017-3305-0.

[8] A. Kumar Jain and B. B. Gupta, “Towards detection of phishing websites on client-side using machine learning based approach,” vol. 68, pp. 687–700, 2018, doi: 10.1007/s11235-017-0414-0.

[9] A. Basit, M. Zafar, A. R. Javed, and Z. Jalil, “A Novel Ensemble Machine Learning Method to Detect Phishing Attack,” Nov. 2020. doi: 10.1109/INMIC50486.2020.9318210.

[10] M. Volkamer, K. Renaud, B. Reinheimer, and A. Kunz, “User experiences of TORPEDO: TOoltip-poweRed Phishing Email DetectiOn,” Computers and Security, vol. 71, pp. 100–113, Nov. 2017, doi: 10.1016/j.cose.2017.02.004.

[11] M. Khonji, Y. Iraqi, and A. Jones, “Phishing detection: A literature survey,” IEEE Communications Surveys and Tutorials, vol. 15, no. 4. pp. 2091–2121, 2013. doi: 10.1109/SURV.2013.032213.00009.

[12] L. Gallo, A. Maiello, A. Botta, and G. Ventre, “2 Years in the anti-phishing group of a large company,” Computers and Security, vol. 105, Jun. 2021, doi: 10.1016/j.cose.2021.102259.

[13] Y. Al-Hamar, H. Kolivand, and A. Al-Hamar, “Phishing attacks in Qatar: A literature review of the problems and solutions,” in Proceedings - International Conference on Developments in eSystems Engineering, DeSE, Oct. 2019, vol. October-2019, pp. 837–842. doi: 10.1109/DeSE.2019.00155.

[14] S. Back and R. T. Guerette, “Cyber Place Management and Crime Prevention: The Effectiveness of Cybersecurity Awareness Training Against Phishing Attacks,” Journal of Contemporary Criminal Justice, 2021, doi: 10.1177/10439862211001628.

[15] F. Roohi and M. Phil, “NEURO FUZZY APPROACH TO DATA CLUSTERING: A FRAMEWORK FOR ANALYSIS,” 2013.