EasyChair Preprint

№ 2461

A Review of Website Phishing Attack Detection Methods

Abdul Basit, Maham Zafar and Zunera Jalil

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

January 26, 2020

A Review of Website Phishing Attack Detection Methods

Abdul Basit[1]
Dept. of Computer Science
Air University
Islamabad, Pakistan
Email: 171266@students.au.edu.pk

Maham Zafar[2]
Dept. of Computer Science
Air University
Islamabad, Pakistan
Email: 171271@students.au.edu.pk

Zunera Jalil[3]
Dept. of Cyber Security
Air University
Islamabad, Pakistan
Email: zunera.jalil@mail.au.edu.pk

Abstract--- Phishing is a type of Cyber-attack that uses fake sites to take sensitive client data, for example, account login certifications, credit card numbers. Phishing sites are commonly entry points of online social engineering attacks, including numerous ongoing on the web scams. In this paper, a review of website phishing attack detection methods discussed the various studies which were used to detect phishing attack. The classification methods, different approaches to detect phishing attack and the obtain results of the studies are discussed briefly.

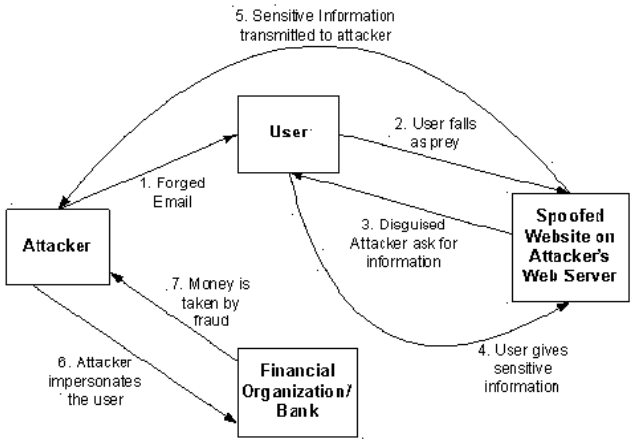
Keywords---Phishing attack, Classification, Data mining, Machine learning

I. Introduction

Phishing sites are commonly entry points of online social engineering attacks, including numerous ongoing on the web scams. In such type of attacks, the attackers create site pages copying genuine sites, and send the suspicious URLs to the targeted victims through spam messages, texts, or online social networking. They will likely target the victim to include their delicate or highly sensitive data (e.g., bank details, government savings number, and so on.).

Despite of the fact that phishing attacks don't require specialized information and these attack procedures are getting comfortable to clients, they are still causing major damages to their financial accounts. These type of attacks likely creates a very negative impact on clients' trust toward the social services such as web services greatly [1].

Phishing is a type of Cyber-attack that uses fake sites to take sensitive client data, for example, account login certifications, credit card numbers, and so on. All through the world, phishing attacks proceed to advance and gain force.



Phishing attack Diagram [2]

In June 2018, the Anti-Phishing Working Group (APWG) detailed upwards of 51,401 special phishing sites. Another report by RSA assessed that worldwide associations endured misfortunes adding up to $9 billion due to phishing happenings in 2016 [3]. These stats have demonstrated that the current anti-phishing arrangements and endeavors are not genuinely effective.

The absolute number of phishing sites recognized by APWG in the 3rd quarter of 2019 was 266,387 [4].

This was up 46 percent from the 182,465 seen in Q2, and practically two fold the 138,328 seen in Q4 2018.

In the 3rd quarter of 2019, APWG part MarkMonitor saw that SaaS and webmail sites remained the greatest focuses of phishing.

Phishers keep on collecting accreditations to those sorts of sites, using them to execute business email comprises (BEC) and to enter corporate SaaS accounts. Stefanie Wood Ellis, Anti-Fraud Product and Marketing Manager at MarkMonitor, noticed: "The top focused on enterprises are to a great extent steady with past quarters."

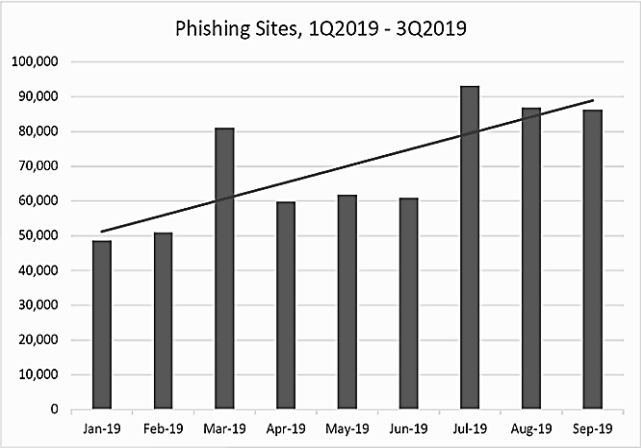


Fig 1. Phishing Activity Trends Report, 3rd Quarter 2019[4]

Phishing can happen in three structures [5]:

- Electronic phishing where a site is copied to take into a website and fools clients into submitting delicate data.
- Email-based phishing, where an attacker sends email to infinite clients assuming some record issue, in trust some of them get victims. Email phishing typically includes electronic phishing also.
- Malware-based phishing where suspicious code is injected into an authentic site and when the client visits that site, the suspicious types of software is introduced on the client's system.

The study has divided in the following sections:

Section (I) contains introduction and how the phishing works and some reports.
Section (II) In this section literature survey is discussed briefly and related work which discussed previous research done regarding the study.
Section (III) we discuss the study and various approaches which were used.
Section (IV) Last section of this paper will contain conclusion.

II. Literature Review

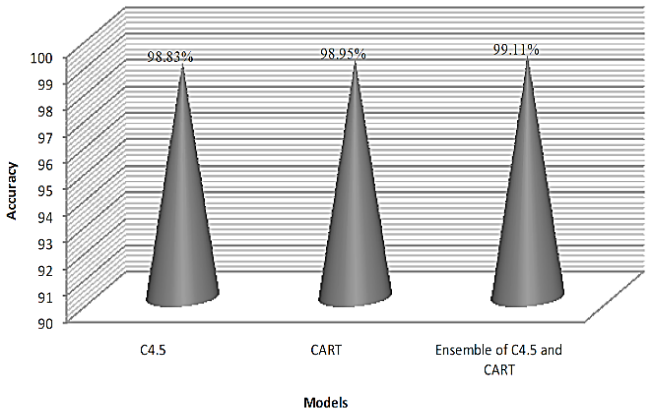
Throughout the decade, numerous strategies have been proposed for detecting phishing website attack. In this section, we will survey few best in class strategies quickly. Following are some of the methods that analysts have used for detecting phishing website attack, which are described below.

- Machine learning based techniques
- Scenario based techniques
- Random forest based techniques
- Hybrid techniques

A. Machine learning based techniques

James, J. [6] taken data set from Alexa and from Phishtank. They used the multiple classifier to predict the accuracy. There proposed approach tell that they got into MATLAB program and read URL one by one and then analyse hostname URL and path these are the feature extraction after feature extraction they evaluate that this is a phishing attack or legitimate activity.
They used four classifiers, NB, DT, K-NN and SVM. Display their data set in to 40% of data set is training and 60% of data set is test data.
They got 93.78% from K-NN after splitting data into 90% in weka. Whereas they Got 91.08% from regression tree in matlab when 60% of data is splitted.
Abdelhamid, N. [7] proposed Machine Learning Comparison based on Models Content and Features and taken a dataset from phishtank and this dataset includes around 11000 examples. They use an approach named eDRI. They claim that dynamic rule induction (eDRI) is the first algorithm of machine learning and deep learning which has been applied on anti-phishing tool.
This algorithm passes data sets with two main threshold frequency and rules strength. The training data set only Store "strong" features these features become the part of rule while other are removed. They used almost 11000 to detect phishing attack.
Mao, J. [8] proposed from page layout feature and taken 49 phishing website dataset from Phishtank.com. Over 20000 texting sample was used in their research. They have used four learning classifiers namely support vector machine (SVM), decision tree, adaboost and random forest.
The result shows that all the classifiers which was used in their study court more than 93% accuracy and more than 84% F1 score. Which shows that their approach is an effective detection in phishing websites. Their study demonstrate that they got enough accuracy but more testing samples can be used for training set.
Kumar Jain, A. [9] proposed and has taken more than two data sets. first from Phish tank which contains 1528 phishing sites, second from Openphish which contain 613 phishing sites, third from Alexa which contains 1600 legitimate sites, forth from payment gateway which contain 66 legitimate sites and fifth from top banking website which contain 252 legitimate sites.
They improved their accuracy by applying machine learning algorithms they use RF, SVM, NN, LR and NB. The maximum accuracy they got from random forest which is 99.09%. They use the feature extraction approach and this feature extraction approach is only on client side.
Hotaa, H. [10] proposed their own feature selection technique and they named as remove replace feature selection technique (RRFST). They claim that they got the phishing email dataset from the khoonji's anti phishing website containing 47 features which day reduced upto only 11 features. The partition of data set is 70% training data set and

30% testing data set. The decision tree was ensemble to predict the performance measures. After calculating the individual accuracies C4.5 classifier got the accuracy of 98.83% and classification and regression tree CART got the accuracy of 98.95%. When both of these ensemble, they got more improved and reached 99.11%.



Accuracy of ensemble model [10]

Anne Ubing, A. [11] proposed their work on ensemble Learning. They used ensemble learning from three techniques which was bagging, boosting, stacking.
Their data set contain 30 features with result column of 5126 record. Data set is taken from UCI machine learning repository which is publicly accessible. They had combine their classifiers to acquire the maximum accuracy. The maximum accuracy which they got from DT which is more than 97%.
CHEN, Y. proposed [12] and used the SMOTE method which improve the detection coverage of the model.
The extracted 28 features which they teach the model and tried a number of data mining techniques including begging, RF, XGB, in result of extracting 28 features from feature evolution method have achieved the higher accuracy through XGB method. They use the data set of Phish tank which has 24471 phishing sites with 3850 legitimate sites.

B. Scenario based techniques

Yao, W. [13] proposed methodology which is mainly includes two processes. One is Logon extraction and the other is identity detection. The proposed methodology tell that the Logon extraction will extract the logo from the image from two dimensional code after performing image processing. And then The Identity detection process will assess the relation between actual identity of the website and it's describe identity if the identity is actual then the website is legitimate website if it is not then this is a phishing website.
They create two data sets which are non-overlapping data set from 726 webpages. The data set contains phishing web pages and legitimate web pages. The legitimate pages are taken from Alexa, whereas the fishing pages are taken from Phishtank.
They believe that logo extraction can be improved in the future. Deep learning technique was not used if deep learning technique or machine learning technique is used then the performance of mobile devices is much improved.
Curtis S.R. [14] proposed and worked on number of persons at its range from 50 to 2885 characters. And they introduce Dark traid attacker's concepts. They used dark traid score to complete the 27 item short dark triad with both attackers and end users were asked to participate in the scenario.
The score based on Psychopathy, Narcissism and Machiavellianism. End-user participant have been very much aware of potential deception keeping in mind the rating for each email their natural work environment will effect.
Williams, E. [15] proposed their study in a workplace. They actually not used any sort of data set. They conduct two studies and both of these studies consider the aspects of emails. The email that is received, the person who received that email and the context in email all the theoretical approaches were studied in that workplace.
They believe that the current study will provide a way to theoretical development in this field. They take 62000 employers over 6 weeks and observe the individuals and targeted phishing emails known as their spear phishing.
Parsons, K. [16] proposed and worked on 985 participants which completed a role which is a scenario based phishing study. They didn't use a dataset because its scenario base research.
It's a two way repeated measures analysis of variance which was named (ANOVA) and it was conducted to assess an effect of email legitimacy and email influence. The email which was used in their research clearly indicates that the recipient has previously donated to some charity. Future work of their study described that they

C. Random forest based techniques

Subasi, A. [17] proposed and used 6 classifiers as which are ANN, K-NN, SVM, RF, RF and C4.5. They discuss in details how these classifiers actually work.
These classifiers are highly accurate in detecting of phishing attack. They used UCI machine learning repository dataset which contains 30 features and 11055 features. They use the WEKA tool to predict the accuracy.
Tyagi, I. [18] proposed and has taken a dataset from UCI machine learning repository which contains 2456 unique URL instances, and a total number of 11055 urls which contains 6157 phishing sites and 4898 are legitimate sites.
There methodology is to input URL and then extract 30 features of URLs and use this features to predict the phishing attack. There are two possible outcomes weather the user has to be notify that the website is phishing or the user has to notify that the website is safe.
They use the machine learning algorithms such as DT, RF, GBM, generalized linear model (GLM) and PCA. They got the maximum accuracy from RF which is 98.40%.
Jagadeesan, S. [19] proposed and uses the method of Random forest (RF) and Support vector machine (SVM). They used

two types of data set the first one is from UCI machine learning repository which has 30 features and One Target feature this data set consists of 2456 entries of phishing and non-phishing urls. Second data set consists of 1353 urls which has 10 features and this URL or categorize in in three classifications. Phishing, non-Phishing and suspicious. They got the maximum accuracy from random forest which is 95.11% on testing data and 94.75% on training data.

Joshi, a. [20] proposed their study on RF algorithm as a binary classifier and reliefF algorithm which is better than any other classifier for feature selection algorithm and it's better than any other combination. The use the forward selection approach which shows the accuracy of 97.63%.

This was done with the 10 features. Whereas 98.13% accuracy was acquired through 48 features. As they can use few more classifiers to acquire the accuracy. But they only used and depend on the RF algorithm and reliefF algorithm. And this is their limitation and gap and this gap will be filled in our research. The dataset which they used was taken from Mendeley website which is publicly accessible. And this data set contain around 10000 and 49 features.

Mao, J. [21] proposed from page layout feature and taken 49 phishing website dataset from Phishtank.com. Over 20000 texting sample was used in their research. They have used four learning classifiers namely SVM, DT, AB and RF.

The result shows that all the classifiers which was used in their study court more than 93% accuracy and more than 84% F1 score. Which shows that their approach is an effective detection in phishing websites. Their study demonstrate that they got enough accuracy but more testing samples can be used for training set.

Koray Sahingoz, O. [22] created their own data set. Which day have uploaded later on a website. The data set contains 73575 urls, and out of this 36400 legitimate URLs and 37175 phishing URLs. As they mentioned that Phishtank doesn't give a free data set on the web page therefore they had created their own data set. They have used 7 different classification algorithms and NLP based features.

The calculated the accuracy and DT got the maximum accuracy which was 97.02%. And the lowest accuracy which they got from AB which was 93.24%. They believe in real time execution.

D. Hybrid techniques

Patil, V. [23] proposed a hybrid solution which will use all three approaches blacklist and whitelist, heuristics and visual similarity. Did all three approaches were used previously but did not used in a hybrid environment. The proposed methodology monitors all traffic on end user system, and compare each URL with the white list which of trusted domains. Website analyse various details for features.

The 3 outcome are suspicious website, phishing website and legitimate website. The machine learning classifier are used to collect data and score is generated. If the score is greater than threshold, then we mark URL as phishing attack and will immediately block it.

They use LR, DT and RF to predict the accuracy of their test websites there are 9076 websites. The highest accuracy got from random forest which is 96.58%.

Niranjan, A. [24] proposed and used the ensembling technique through voting and stacking method. And they reduced the features. The data set is taken from UCI machine learning phishing data set they remove the relevant features and take only 23 features out of 30 features. And tried to improve their accuracy. T

he data set contain 6157 legitimate and and 4898 phishing instances out of a total of 11055 instances. They used EKRV model which is a hybrid technique which involves the combination of K-NN and RC which day combine for voting for stacking.

Leng Chiew, K. [25] proposed and used 5000 phishing web pages based on URLs from Pishtank and OpenPhish. Another 5000 legitimate web pages based on URLs from Alexa and the common Crawl5 archive. Basically they ensemble the hybrid strategy there was two major types of ensemble techniques, namely data perturbation and function probation. They used six classifiers and random forest was the major classifier which was compare by its full features and baseline features.

Pandey, A proposed [26] and used the data-set of repository of the University of California. The dataset has 10 attributes and 1353 instances. Use the train RF and SVM hybrid model which they utilize to predict the accuracy.

TABLE I. Comparative Study of Previous Work

Authors	Classification	Feature selection technique	Accuracy rate
James, J [5]	J48,IBK,SVM, NB	-	89.75%
Subasi, A [17]	ANN, kNN, RF, SVM, C4.5, RF	-	97.36%
Abdelhamid, N [6]	eDRI	-	93.5%
Mao, J. [35]	SVM, DT	-	93%
Kumar Jain, A [8]	-	Feature extraction	99.09%
Yao, W [13]	-	Logo Extraction	98.3%
Patil, V [24]	LR, DT, RF	-	96.58%
Jagadeesan, S.[32]	RF,SVM	-	95.11%
Hotaa, H.[9]	CART, C4.5	RRFST	99.11%
Tyagi, I [19]	DT, RF, GBM	PCA	98.40%
Curtis S.R. [14]	-	-	-
Koray Sahingoz [23]	SVM, DT, RF, kNN, KS, NB	NLP	97.98%
Parsons, K. [37]	-	-	-
Joshi, a [21]	RF, RA	RA	97.63%
Anne Ubing [11]	EL	-	95.4%
Mao, J [22]	SVM, RF, DT, AB	-	97.31%
Williams, E. [31]	-	-	-
Niranjan, A [25]	RC, kNN, IBK, LR, PART,	-	97.3%
CHEN, Y. [12]	ELM, SVM, LR, C4.5, LC-ELM, kNN, XGB	ANOVA	99.2%
Leng Chiew, K [26]	RF, C4.5, PART, SVM, NB	-	96.17%
Pandey, A [27]	SVM, RF	-	94%

TABLE II. Acronyms lists used in Table I

SVM:	Support Vector Machine
RF:	Random Forest
IBK:	Instant Base Learner
ANN:	Artificial Neural Network
RF:	Rotation Forest
DT:	Decision Forest
J48:	C4.5
eDRI:	Enhanced Dynamic Rule Induction
LR:	Linear Regression
CART:	Classification and Regression tree
XGB:	Extreme Gradient Boost
GBDT:	Gradient boosting decision tree
AB:	AdaBoost
GBM:	Gradient Boosting Machine
NB:	Navies Bayes
kNN:	K-Nearest Neighbor
KS:	K-star
LC-ELM:	Combination Extreme Learning Machine
ELM:	Extreme Learning Machine
RC:	Random Committee
PCA:	Principle component analysis

III. Discussion

A comparative study of previous works which have used different approaches as discussed above were machine learning approach, scenario based approach and classification of random forest approach and hybrid techniques were used. These all approaches were most used approaches to detect website phishing attack.

The machine learning methods are the most common and effective methods to detect phishing attack. Different classification method is used such as SVM, RF, ANN, C4.5, k-NN, PCA, DT which are the most effective way to detect phishing attack.

The most classification method to detect website phishing attack were random forest as it got the highest accuracy among any other classification methods. The random forest classification method is used on different datasets and got the highest accuracy among other classification methods. The various studies proved that the got more than 95% accuracy with classification of random forest. The common dataset which is used by researchers is UCI machine learning dataset which have 11055 instances.

In various studies the researchers create a scenario based environment to detect phishing attack but this is not useful in all environments because every organization individual has different behavior and most of individuals will be aware of the scenario. They believe some of scenario based approaches are logo extraction which can be improved in the future.

The hybrid model is another way to detect phishing attack as it got sometimes more accuracy than that of random forest. To get a highest accuracy to detect phishing attack the hybrid model is used. They believe that by ensembling we can get a highest accuracy than any other methods. They used EKRV model which is a hybrid technique which involves the combination of K-NN and RC which day combine for voting for stacking.

IV. Conclusion

As malicious URLs are created day by day easily so attackers can create a technique to fool the users and modify the URLs to appear legitimate to attack. once the modified URLs appear to be the legitimate ones the attack launches.

Nowadays deep learning and machine learning methods are used to detect phishing attack. classification methods such as RF, SVM, C4.5, DT, PCA, k-NN are most common. These methods are most useful and effective for detecting the phishing attack. Future direction includes more scalable and robust method including feature reduction to detect theses phishing attacks.

REFERENCES

[1] (2019). Retrieved 21 November 2019, from http://docs.apwg.org/reports/apwg_trends_report_q4_2016.pdf

[2] (2020). Retrieved 20 January 2020, from https://www.researchgate.net/publication/235947501_Analysis_of_Phishing_Attacks_and_Countermeasures/figures?lo=1

[3] Forecast, 2. (2019). 2017 Global Fraud and Cybercrime Forecast. from https://www.rsa.com/en-us/blog/2016-12/2017-global-fraud-cybercrime-forecast

[4] (2019). Retrieved 27 November 2019, from https://docs.apwg.org/reports/apwg_trends_report_q3_2019.pdf

[5] Dong, Z. (2015). "Beyond the lock icon: real-time detection of phishing websites using public key certificates". *IEEE*.

[6] James, J. (2013). "Detection of Phishing URLs Using Machine Learning Techniques". *In 2013 International Conference on Control Communication and Computing (ICCC).* IEEE

[7] Abdelhamid, N. (2017). "Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features". *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*.

[8] Mao, J. (2019). "Phishing page detection via learning classifiers from page layout feature". *EURASIP Journal on Wireless Communications and Networking*.

[9] Kumar Jain, A. (2017). "Towards detection of phishing websites on client-side using machine learning based approach". *Telecommunication Systems*.

[10] Hotaa, H. (2019). "An Ensemble Model for Detecting Phishing Attack with Proposed Remove-Replace Feature Selection Technique". *International Conference on Computational Intelligence and Data Science (ICCIDS 2018).* Elsevier.

[11] Anne Ubing, A. (2019). "Phishing Website Detection: An Improved Accuracy through Feature Selection and Ensemble Learning". (*IJACSA) International Journal of Advanced Computer Science and Applications, 10(1).*

[12] CHEN, Y. (2019). "Machine Learning Mechanisms for Cyber-Phishing Attack". *IEICE TRANS. INF. & SYST*, E102–D.

[13] Yao, W. (2018). "LogoPhish: A New Two-dimensional Code Phishing Attack Detection Method". *2018 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing,*

Social Computing & Networking, Sustainable Computing & Communications.
[14] Curtis S.R. (2018) "Phishing attempts among the dark triad: Patterns of attack and vulnerability", *Computers in Human Behavior (2018),*
[15] Williams, E. (2019). Exploring susceptibility to phishing in the workplace. International Journal of Human-Computer Studies, 120, 1-13
[16] Parsons, K. (2019). Predicting Susceptibility to Social Inßuence in Phishing Emails. International Journal of Human-Computer Studies.
[17] Subasi, A. (2017). "Intelligent Phishing Website Detection using Random Forest Classifier". *2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA).*
[18] Tyagi, I. (2018). "A Novel Machine Learning Approach to Detect Phishing Websites". *2018 5th International Conference on Signal Processing and Integrated Networks (SPIN).* IEEE.
[19] Jagadeesan, S. (2018). URL Phishing Analysis using Random Forest. International Journal of Pure and Applied Mathematics, 118(20).
[20] Joshi, a. (2019). "Phishing Attack Detection using Feature Selection Techniques". *International Conference on Communication and Information Processing (ICCIP-2019).* Elsevier.
[21] Mao, J. (2019). "Phishing page detection via learning classifiers from page layout feature". *EURASIP Journal On Wireless Communications and Networking*.
[22] Koray Sahingoz, O. (2019). "Machine learning based phishing detection from URLs". *Expert Systems with Applications.*
[23] Patil, V. (2018). "Detection and Prevention of Phishing Websites using Machine Learning Approach". *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA).* IEEE.
[24] Niranjan, A. (2019). "EKRV: Ensemble of kNN and Random Committee Using Voting for Efficient Classification of Phishing. Progress" *Advanced Computing and Intelligent Engineering, Advances in Intelligent Systems and Computing*
[25] Abdelhamid, N. (2017). "Phishing Detection: A Recent Intelligent Machine Learning Comparison based on Models Content and Features". *2017 IEEE International Conference on Intelligence and Security Informatics (ISI).* IEEE.
[26] Pandey, A. (2018). "Identification of Phishing Attack in Websites Using Random Forest-SVM Hybrid Model". *ISDA.* Springer