IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115128

Phishing Attack Detection using Hybrid Learning

Shreetej Sharma[1], Darshan M[1], Shashank KS[1], Prof. Usha C.R[2]

[1]Student, Department of Information Science & Engineering, Global Academy of Technology, Bengaluru, India

[2]Assistant Professor – Department of Information Science & Engineering,

Global Academy of Technology, Bengaluru, India

Abstract: In these tough times, during a pandemic, when a virus of the size of few microns has taken the whole world back. With the global economy declining, people are finding new ways to make money online, sometimes also through illegal means. Phishing is regarded as one of the most dangerous risks to internet users, and it is growing at an exponential rate. As a result, hackers have become more innovative in their assaults and have been able to execute them on a big scale. A phishing attack works by creating an accurate clone of an actual site and directing people to the site's page. Because the site's page is deceptive and identical to the actual, legitimate individuals are frequently duped into performing activities on such pages. Phishing is a type of assault that combines foundations of social engineering with emerging technological approaches. This wrapping of the faux site to appear to be the real one persuades the user to give their identity. Our system developed with the concepts of Hybrid learning which is a amalgamation of Machine Learning and Deep Learning, aims to detect such Phishing attacks by marking websites as legitimate or phishy.

Keywords: AWPG, Hybrid Learning, Link Guard Algorithm, Phishing attack, Website,

1. INTRODUCTION

Phishing is a fraudulent activity in which the phisher/attacker attempts to manipulate internet users into revealing personal information/credentials in attempt to profit handsomely [1]. Phishing is similar to fishing, but the purpose is different; in phishing, the attacker utilises bait (sending an email with an embedded hyperlink that goes to a malicious site) to acquire internet users' credentials. Previously, hackers were known as Phreaks (a Phreak is someone who unlawfully sneaks into telephone networks to make free long-distance phone calls or tap phone lines) and are closely tied to one another. The substitution of "f" for "ph" is intended to associate phishing schemes with phreaks [2, 3]. For the past two decades, phishing has been the most severe assault, and there are several attacks every day [4–6]. The first phishing fraud was discovered on American online (AOL), a provider of internet services, on January 2, 1996 [2]. The phisher produces credit card numbers at random and uses those credit card numbers to establish AOL accounts.

Later, using AOL instant messengers or email, they send an email to clients requesting them to verify their account information by clicking on the embedded URL supplied in the email. If the user clicks on the URL and enters their credentials, the attacker receives this information automatically. As a result, the attackers utilise such credentials for fraudulent purposes.

During the pandemic, Phishing attacks reached all time high, the AWPG activity trends report shows that Phishing Reaches Monthly Record in Q3; Attacks Doubled since Early 2020[7]. With people locked in their houses wanting to make extra income, some of them were duped to enter confidential details on phishy sites, which ultimately resulted in monetary loss. To overcome this problem, different approaches for detecting such phishing attacks have been developed; we will explore some of these methodologies in this paper.

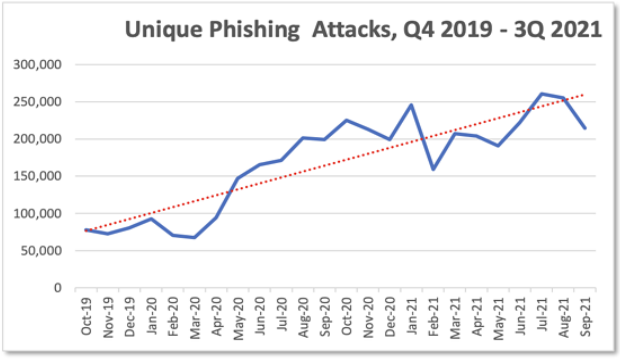


Figure 1. Phishing Activity Trends Report for Q3 2021

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115128

The following is how this paper is structured. We continue form this point with the literature survey that acts as a background study for Phishing Attacks and their detection using alternate methodologies. The findings are then provided in a tabular form to highlight the study's findings. The purpose of this study is to review certain major literatures that are required for system implementation.

2. LITERATURE SURVEY

Phishing is an appalling threat within the internet security domain. In this attack, the user inputs his/her personal data to a dupe web site that appears like a legitimate one. we have conferred a survey on phishing detection approaches supported by visual similarity. This survey provides a much better understanding of phishing web site, various solution, and future scope in phishing detection. several approaches are mentioned during this paper for phishing detection; but, most of the approaches still have limitations like accuracy, the measure against new phishing websites, failing to observe embedded objects, and then forth. These approaches use varied attributes of a webpage to observe phishing attacks, like text similarity, font color, font size, and pictures present within the webpage. Text primarily based similarity approaches are comparatively quick, however they are unable to espy phishing attack if the text is replaced with some image. Image processing-based approaches have high accuracy rate whereas they are complicated in nature and are long. moreover, most of the work is completed offline. These involve data assortment and profile-creation phases to be completed initially. A comparative table is prepared for simple glancing at the advantages of the assorted offered approaches developed by totally different individuals and published. No single technique is enough for adopting it for phishing detection functions. Detection of phishing websites with high accuracy continues to be an open challenge for additional analysis and development.

Table 1. Table of Comparative Analysis of the Literature Survey

Sl. No	Paper title, Author	Year of Publication	Methodology / Algorithms used	Results obtained
1	Detecting Phishing-Sites using Hybrid Model, Poonam Kumari, Apoorva H R Gowda, Bhandhavya K, Bhavya M U and Spurthi M N	2020	UCI repository- Random Forest, Decision Tree, Sequential Minimal Optimization, Bayesian net, Naïve Bayes, Fuzzy Unordered Rule and Instance based Learning	Deep research and analysis resulted in union of classification model as IBK which gave better results in comparison with individual model in terms of enhanced accuracy.
2	Hybrid Machine Learning: A Tool to Detect Phishing Attacks in Communication Networks, Ademola Philip Abidoye and Boniface Kabaso	2020	PhishTank- Support Vector Machines (SVMs) Classifiers, Naïve Bayes Classifiers	Conducted experiments showed better performance by achieving a highest classification accuracy with a low false-positive rate of 1.06%.
3	A Hybrid Approach for Phishing Website Detection Using Machine Learning, Harsh Kansagara, Vandan Raval, Faiz Shaikh, Prof. Saniket Kudoo	2021	Random Forest Algorithm, TF-IDF approach	The system made a safe environment for browsing by detecting (with high accuracy) phishing websites keeping the use safe.
4	A Hybrid Machine Learning based Phishing Website Detection Technique through Dimensionality Reduction, Nusrath Tabassum, Farhin Faiza Neha, Md. Shohrab Hossain, and Husnu S. Narman	2021	Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), XGBoost and several hybrid classifiers such as RF + XGBoost, DT + XGBoost, DT + RF, DT + RF + XGBoost, SVM + DT +	Robust feature selection techniques resulted in highest accuracy (98.2%) which was done by reducing the dimensionality of feature subset.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115128

			XGBoost,SVM + DT + RF, LR + DT + RF + XGBoost	
5	Phishing detection system using machine learning classifiers, Nur Sholihah Zaini, Deris Stiawan, Mohd Faizal Ab Razak, Ahmad Firdaus, Wan Isni Sofiah Wan Din, Shahreen Kasim, Tole Sutikno	2020	WEKA tools, Random Forest, J48, Multi-Layer Perceptron (MLP) and K-Nearest Neighbors (KNN)	Random forest classifiers have achieved the highest accuracy result of 94.79 percent when compared to KNN which achieved only 93. 08 percent.
6	A Machine Learning Approach for Phishing and Its Detection Techniques, Dhananjay Merat, Anurag Patil, Sourabh Gavsane, Vivekanand Jadhav, Prof. Himanshu Joshi	2020	Support Vector Machines and Naïve Bayes Classifier	The different methods that have achieved highest accuracy have been tested by selecting minimum number of features and by reducing the dimensionality of feature subset.
7	Hybrid Rule-Based Model for Phishing URLs Detection, Kayode S. Adewole (&), Abimbola G. Akintola, Shakirat A. Salihu1, Nasir Faruk, and Rasheed G. Jimoh	2019	Rule Induction Algorithm	All experiments were conducted using R statistical package and Rweka library.
8	Detection of Phishing Websites Using Hybrid Model Ch. Chakradhara Rao, A. V. Ramana	2018	Decision Tree, IBK, Naïve Bayes and Bayes Net Algorithms	Experiments have been done to measure the accuracy of all the algorithm at the beginning, since the accuracy measure of Naïve Bayes algorithm is very low when compared to other algorithm such as Random Forest, IBK and Decision Tree.
9	Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages Yun Lin and Ruofan Liu,; Dinil Mon Divakaran, Trustwave; Jun Yang Ng and Qing Zhou Chan, N Yiwen Lu, Yuxuan Si, and Fan Zhang,; Jin Song Dong,	2021	Object detection Model and Siamese model	All discovered phishing webpages and their reports are published and are, compared to other baseline approaches
10	Hybrid Model of Phishing Email Detection: A Combination of Technical and Non-Technical Anti-Phishing	2018	Proposed Anti Phishing Models	Separated evaluation processes have been conducted to individually evaluate each of the modules from which the system is consisting of.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

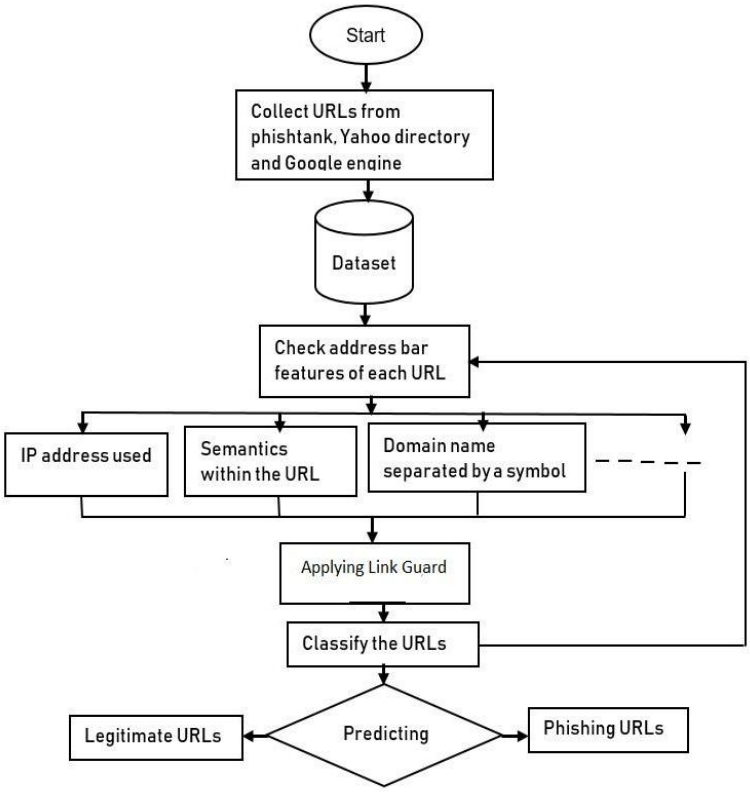
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115128

	Approaches Melad Mohamed Al-Daeef, Nurlida Basir, and Madihah Mohd Saudi			

3. METHODOLOGY

The Proposed Framework for Detecting Phishing URLs.



The approach is divided into two parts, and each part's output is an input to the next part as shown in the proposed framework.

The first part is based on data collection, processing of data sets, and URLs feature extraction. We consider different heuristic features in the structure of URLs, ranging from a generic social engineering feature, lexical feature in the URL, multiple alphabets, and phishing target brand name. The feature vector is constructed with important features to model our classifiers. The second part is based on the classification of data set using link guard to evaluate our approach.

An end-host based anti-phishing algorithm which we call Link Guard, based on the characteristics of the phishing hyperlink. Since Link Guard is character-based, it can detect and prevent not only known phishing attacks but also unknown ones. Link Guard is light-weighted in that it consumes very little memory and CPU circles, and most importantly, it is very effective in detecting phishing attacks with minimal false negatives. Link Guard detects 195 attacks out of the 203 phishing archives provided by APWG without knowing any signatures of the attacks.

A. Processing of Data Sets and URLs Features Extraction

A large number of data sets collected and processed to make them suitable for the requirement of this study. The processing involved many stages, these include webpages feature extraction, data standardization, and attribute weighing. These steps are very important so that the classifiers would be able to understand the data sets and appropriately categorize them into their classes. The classifier is regularly trained with new phishing web pages to learn new trends in phishing. The outcome of this phase is used as input to the next part of the Link guard classifiers.

We propose a hybrid machine learning approach to effectively classify phishing URLs based on the information available to an individual URL. Phishing URLs are treated as a binary classification problem with the benign URLs belong to the

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 5, May 2022

DOI: 10.17148/IJARCCE.2022.115128

negative class and phishing URLs belonging to the positive class. Phishing URLs are collected from Phish Tank, Yahoo directory, and the Google engine to form our data sets. Thereafter, it extracts many features that have proved to be effective in predicting phishing URLs by employing different publicly available resources to classify the data sets into their respective classes. We apply Link Guard Algorithm to create models from training data sets which consist of feature extractions and class labels.

B. Phishing Data Sets

Phish Tank is a joint project to which people can submit suspicious phishing URLs for confirmation. It is a public clearinghouse for phishing URLs . Suspicious URLs are further scrutinized by many people before being confirmed as phishing URLs and added to a blacklist. Phish Tank provides a comprehensive list of current and active phishing URLs. Researchers and developers can download phishing URLs from the Phishing Web site after signing up. They would be able to download the URLs from Phish Tank in different file formats with an API key. Also, System observe that phishers constantly develop new tactics to get personal information from unsuspecting users, to explore various and recent methods the attackers are using. Phishers also use this period as an opportunity to display their tactics and launch different attacks on unsuspecting users.

C. Legitimate Data Sets

The URLs are collected from the Yahoo directory. Yahoo provides a generator that arbitrarily produces an URL in its directory each time the Web page is visited. This service is used to randomly choose an URL and download the contents of the Web page with the server header information. Our list consists of URLs from financial institutions, e-commerce, online services, cloud storage, religious organizations to get different URL structures and Web page contents. To provide more learning instances for legitimate URLs. DMOZ is a multilingual open-content directory of World Wide Web links containing more than three million URLs.

We use a Google tool to analyze the list of URLs collected and crawled. These URLs are used as legitimate webpages based on the assumption that all the URLs extracted were benign since they were downloaded from legitimate Internet sources.

Python is used to parse the legitimate and phishing URLs and extract the features discussed. Web pages that we could not extract features from their contents were discarded to get only valid URLs for our data sets.

D. Data Authentication

Data sets collected need to be authenticated to ascertain the real status of the URLs, particularly in the case of phishing websites as it is known that the phishing website only lasts a few weeks. Thus, every URL needs to be authenticated before processing.

In this section, we present relevant features that are effective in predicting phishing web sites. Each feature is discussed with its associated rules.

3.1 A generic social engineering feature

Phishers use generic greetings in their messages such as “Sir”, “Dear Bank Customer”, “Dear Customer”, and “Dear Member” to address their target victims. The content of the message is always threatening such as “please update your bank account to prevent it from being blocked”, "Your account has been compromised!", "Urgent action required!", "Your account will be closed!" These intimidation strategies are becoming more common than the promise of "instant riches"; taking advantage of victims’ anxiety and concern to get them to provide their personal information.

Rule: $\boldsymbol{if}$ *the greeting is directed to account owner and do not require to supply* a piece of *personal information via a link in the message*→ *Legitimate*

$\boldsymbol{else\ if}$ the *greeting is generic*→ *Suspicious*

$\boldsymbol{else}$ *update your information via a given link*→ *Phishing*

3.2 IP-based URL

Internet Protocol (IP) address is one of the ways to hide the webpage address. If an IP address is used instead of a Domain Name System (DNS) address in the URL, it will be difficult for innocent users to ascertain where they are being directed to when they click the link or press the Enter key on their system to load the page. Another reason for using the IP address is that phishers would not like to spend money to buy a domain for their phony web pages.

Rule: $\boldsymbol{If}$ *the domain name has an IP Address* → *Phishing*

$\boldsymbol{else}$→*Legitimate*

3.3 Long URL to hide the fake part

Attackers can use lengthy URLs to mask the fake part in the address bar. For instance,

DOI: 10.17148/IJARCCE.2022.115128

"http://prudentbank.com/2k/ab51e2e319e51502f416dbe46b773a5e/?cmd=_home&dispatch=11004d58f5b74f8dc1e7c2e8dd4105e811004d58f5b74f8dc1e7c2e8dd4105e8@phishing.net.html"

We computed the length of URLs in our data sets and determined their average length to ensure the accuracy of our research. The findings showed that if the URL length is less than 52 characters, it is classified as legitimate; it is suspicious if the length is between 52 and 73 characters, and it is a phishing URL if the URL is more than 73 characters. A method based on frequency has been used to update this feature rule, which improves its accuracy.

Rule: $\boldsymbol{If}\ URL\ length<52\ characters \rightarrow Legitimate$

$\boldsymbol{else\ if}\ URL\ length \geq 52\ and \leq 73\ characters \rightarrow Suspicious$

$\boldsymbol{else} \rightarrow Phishing$

3.4 Shortened URL "Tiny URL"

Short URL enables to reduce long links from social networks and top sites on the Internet. This is achieved by the service provider through an "HTTP Redirect" on a domain name that is short and redirects to the corresponding long URL [17]. For instance, an URL for Wiki's article "http://en.wikipedia.org/wiki/URL shortening" contains 64 characters and its corresponding short URL http://bit.ly/c1htE; it contains 16 characters with Bitly's default domain name "bit.ly" and the hash "c1htE" as the back-half. A hash only consists of letters and numbers "a-z, A-Z,0- 9". Attackers use this shortened URL feature to hide links to infected websites or phishing.

Rule: $\boldsymbol{if}\ TinyURL \rightarrow Phishing$

$\boldsymbol{else} \rightarrow Legitimate$

4. CONCLUSION

In this paper, we addressed the term phishing and gave a thorough classification of phishing attacks depending on the goal of getting crucial data of victims. This paper presents a comprehensive study of phishing attacks as well as the many methods for carrying out phishing attacks (i.e., through email, advertisements, Instant messaging, phone calls, social media sites, malware, website, DNS etc.).

There are also some real-time phishing attacks that correlate to each sort of attack listed in the paper. The attack information retrieved from APWG survey reports and OpenPhish phishing feeds is used for statistical analysis of phishing attacks. According to the data, China is the most impacted country by phishing, while America is the top host of phishing. Payment sectors are the most attacked by phishing and Trojans are the most commonly used malware for phishing. We also found that over 45% of phishing URLs are HTTPS protected.

This paper also presents, a new classification technique using Link Guard Algorithm, which is quite effective in determining a web-link's legitimacy and alerts the user to prevent any accidental data breach by the user. It also ensures safe user presence over the internet by shielding users from malicious or unsolicited links in Web pages and Instant messages.

REFERENCES:

1. Kirda E, Kruegel C. Protecting users against phishing attacks with antiphish. In annual international computer software and applications conference 2005 (pp. 517-24). IEEE.
2. http://www.phishing.org/history-of-phishing. Accessed 10 May, 2022.
3. Mei Y. Anti-phishing system: detecting phishing e-mail. School of Mathematics and Systems Engineering. 2008.
4. https://dictionary.cambridge.org/dictionary/english/phishing. Accessed 8 May, 2022.
5. Yadav S, Bohra B. A review on recent phishing attacks in internet. In international conference on green computing and internet of things 2015 (pp. 1312-5). IEEE.
6. S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs, "Who falls for phish? a demographic analysis of phishing susceptibility and effectiveness of interventions," in Proceedings of the 28th international conference on Human factors in computing systems, CHI '10. New York, NY, USA: ACM, 2010, pp. 373–382.
7. APWG Q1-Q3 Report, 2015, http://docs.apwg.org/reports/apwg trends report q1-q3 2015.pdf
8. B. Parmar, "Protecting against spear-phishing," Computer Fraud & Security, vol. 2012, no. 1, pp. 8–11, 2012.
9. W. Jingguo, T. Herath, C. Rui, A. Vishwanath, and H. R. Rao, "Phishing susceptibility: an investigation into the processing of a targeted spear phishing e-mail," IEEE Transactions on Professional Communication, vol. 55, no. 4, pp. 345–362, 2012.
10. T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer, "Social phishing," Communications of the ACM, vol. 50, no. 10, pp. 94–100, 2007.
11. C. H. Hsu, P. Wang, and S. Pu, "Identify fixed-path phishing attack by STC," in Proceedings of the 8th Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference (CEAS'11), pp. 172–175, ACM, Perth, Australia, September 2011.

DOI: 10.17148/IJARCCE.2022.115128

12. N. A. G. Arachchilage and M. Cole, "Designing a mobile game for home computer users to protect against phishing attacks," https://arxiv.org/abs/1602.03929.
13. https://www.cisco.com/c/en_in/products/security/email-security/what-isphishing.html
14. Amaad Ul Haq Tahir, Sohail Asghar, Ayesha Zafar,: A Hybrid Model to Detect Phishing Sites using Supervised Learning Algorithms, In: 2016 International Conference on Computational Science and Computational Intelligence (CSCI), 15-17 Dec. 2016
15. Poonam Kumari, Apoorva HR Gowda, Bhandhavya K: Detecting Phishing-Sites using Hybrid Model, In: International Journal of Engineering Research & Technology, 2020
16. Ademola Philip Abidoye, Boniface Kabaso,: Hybrid Machine Learning: A Tool to Detect Phishing Attacks in Communication Networks, In: International Journal of Advanced Computer Science and Applications, Vol. 11, No. 6, 2020
17. https://docs.apwg.org/reports/apwg_trends_report_q3_2021.pdf
Phishing Activity Trends Report, 3rd Quarter 2021