**Analysis of Machine Learning Techniques for Detecting and Analyzing Phishing Websites**

***Yogi Reddy M1, Pagalla Bhavani Shankar2***

***1Assistant Professor, Department of Computer Science & Engineering, GITAM Deemed to be University, Telangana, India, ORCID:******0000-0002-2102-1869***

***2Assistant Professor, Department of Computer Science & Engineering, University College of Engineering and Technology, Krishna University, Machilipatnam, India, ORCID:0000-0001-5935-7758***

**Abstract:** Phishing is a technique of social engineering used by hackers to steal personal information like passwords and credit card numbers. Phishing domains must be found in order to stop attacks from happening. While there is no foolproof way to spot phishing sites. Identifying phishing attempts is crucial here. Many different approaches to solving this problem, and in particular how to identify phishing websites, have been proposed in the literature. This paper discusses the primary obstacles and findings in this area, providing a thorough and comprehensive evaluation of the state of the art. In this research, we offer a method that makes use of machine learning models that have been trained to spot phishing domains. Furthermore, it provides a comparison of the various machine learning techniques now in use for the study and categorization of phishing and legitimate websites, as well as a full examination of the assaults themselves.

1. **Introduction:**

Phishing attack identification and prevention are collectively referred to as "fishing detection," "fraud detection," or "anti-phishing." By masquerading as a reputable company or institution, phishers attempt to deceive their victims into divulging important information such as login credentials or financial data. The need to safeguard individuals and businesses from falling prey to phishing schemes is what makes fishing detection so crucial. Consequences of falling for these scams include but are not limited to, personal information theft, monetary loss, and security breaches. People and businesses can protect themselves from phishing attacks by using fishing detection methods. Among the many methods used to uncover fishermen at work are:

Automatic detection and blocking of malicious emails and websites commonly used in phishing attacks can be achieved through the use of filters and security mechanisms built into email systems and web browsers. Sender reputation, symptoms of phishing, and harmful links are only some of the factors that these filters examine.

URL analysis is commonly used by fishing detection systems to verify the reliability of a link. They search for red flags like misspellings, weird domain names, and unexpected links. With this study, phishing websites that are set up to trick consumers can be located and avoided.

Phishing detection systems analyze the content of emails, messages, and websites using machine learning and natural language processing techniques. They are on the lookout for indicators of harmful intent, such as certain patterns, buzzwords, and so on. Emails with urgent requests, spelling mistakes, or strange layouts could raise suspicions of phishing.

Monitoring the reputation of domains and adding them to blacklists is an effective method for identifying and blocking harmful websites. New phishing operations can be thwarted in real time thanks to these databases, which are regularly updated with information about reported events.

Providing users with awareness training and information on phishing dangers can greatly lessen the likelihood that they will fall for a phishing scam. Users become adept at spotting telltale signs of phishing attempts, such as questionable emails from unknown senders or requests for personal information. Phishing attacks are frequently simulated in training programs to help users in spotting and report suspicious behavior.

Organizations may keep up with the most recent phishing methods and attack vectors using real-time threat intelligence by working with threat intelligence providers and keeping tabs on emerging phishing trends. With this knowledge, we can create better defenses and preventative actions.

Keeping up with the ever-evolving phishing schemes and assault methods is a constant challenge when it comes to fishing detection. Thus, to effectively prevent phishing attacks, a combination of technology solutions, user awareness, and continual monitoring and adaption is required. Individuals and businesses can strengthen their cybersecurity defenses and reduce their vulnerability to phishing attempts by instituting fishing detection mechanisms.

**1.1 Deceptive phishing attack**

Phishing is a sort of online identity theft in which sensitive information is stolen by sending bogus e-mails to unwary users, who are then directed to fake websites. User names and passwords, social security numbers, credit card numbers, bank account numbers, and even more intimate details like dates of birth and mothers' maiden names are all fair game for phishers. Most attacks are quite simple, like asking a victim for his bank account information and PIN, but they are also extremely effective. There is a consistent informational pattern shared by all phishing scams. Fig.1 depicts a typical phishing attack flow, with each stage shown as follows:

1. The phisher created a phishing website to target users.
2. The phisher sends out a flood of fake emails with a "call to action" that encourages the receiver to visit a malicious website.
3. The recipient is tricked into visiting the phishing website by the counterfeit email, at which point her personal information is at risk.
4. The information is sent to the phisher from a phishing server.
5. The phisher impersonates the victim on the compromised website in order to get their financial information.



Fig.1 depicts a typical phishing attack flow

**1.2 Detecting a fishing attack**

Because phishers constantly improve their methods, it might be difficult to spot their fraudulent emails and websites. However, there are a number of warning indicators and preventative steps you can take to avoid falling victim to a phishing scam:

* Verify the Email Address of the Sender: Check the email's sender's address to make sure it's legit. Email addresses used by phishers sometimes look like those used by legitimate businesses but with typos or other minor changes. Be wary of messages coming from unknown senders or domains.
* Check the Email's Content: Scam emails frequently have typos, misspellings, and bad formatting. Keep an eye out for requests for information that seem too good to be true, too urgent, or too personal. Communication within legitimate organizations is typically polished and error-free.
* Clicking on a link in an email or message without first verifying it with a mouse hover is a bad idea. It is important to double-check that the status bar or tooltip URL corresponds to the desired location. URLs that contain misspellings, extra letters, or a different domain name should be used with caution.
* Be wary of any urgent or threatening language, as this is a common tactic used by phishers to get their targets to act quickly. They might threaten to close your account, inform you that you've won a prize, or warn you of a potential security breach. Authentic businesses tend to speak in a more measured, businesslike tone.
* The use of HTTPS protocol and a current SSL certificate are signs of a legitimate website, and should be verified before disclosing any personal information. Check that the website's address begins with "https://" rather than merely "http://" and that the padlock icon appears in your browser's address bar.
* Be wary of anyone who contacts you via email or other means asking for personal, financial, or login information. Email requests for this type of information from legitimate businesses are quite unusual. If you have any questions, please do not hesitate to contact the company through email or phone.
* Put Anti-Phishing Software and Other Tools to Use: Protect yourself from phishing by using anti-virus software, browser add-ons, or spam filters. Using these resources, you can spot and stop common phishing scams.

Keep in mind that there is no failsafe method of detection because hackers are always changing their methods. It is safer to alert the appropriate company or your IT department if you even believe a phishing attempt has been made.

1. **State-of-the-art techniques for detecting phishing websites**

In this part, you'll find a comprehensive and up-to-date literature evaluation that backs up your research questions and their potential answers. Phishing has evolved over the years, with perpetrators using ever-more-creative methods by carefully monitoring their victims and employing cutting-edge tools to make their fake websites look more trustworthy than ever before [2]. The chapter provides an overview of phishing and analyses the relevant proposed method. The low detection accuracy and high false alarm rates of existing phishing detection methods are depicted in Fig 2 along with their general overview.



Fig 2: Methods for detecting phishing attacks: an overview.

Even though blacklists are the most frequent tactic, they are useless in stopping phishing assaults since it is so simple to create a new domain, and even the most comprehensive blacklist can't guarantee a perfectly up-to-date strategy [3]. Bhawna Sharma and Parvinder Singh's [4] hybrid method, which they name PhishAlert, combines content similarity whitelists, style similarity, and heuristics. PhishAlert performed better than two existing algorithms, CANTINA and CANTINA+, on an experimental dataset of 500 phishing sites and 500 legitimate sites [4]. Both the legitimate and phishing URLs have been retrieved; the former came from the stuffgate server [5] and the latter from phishtank [6].

Fig 3: Comparison of CANTINA and CANTINA+ vs PhishAlert in Terms of Accuracy

As can be shown in Fig 3, PhishAlert has been proven to be 98% accurate. PhishAlert's effectiveness declines as datasets get bigger. The authors suggested adding more features to future iterations of the PhishAlert algorithm to improve detection efficiency.

Using Resource Description Framework (RDF) and Random Forests, Vamsee Muppavarapu, Archanaa Rajendran, and Shriram Vasudevan suggested [7] a method for detecting phishing attempts. The purpose of this work is to catalogue known phishing domains and suggest new ones. There are two stages to the process. Both the first and second stages are based on machine learning techniques, with the first stage based on the RDD model of web pages. Together, the two tiers work to improve the system's accuracy and precision by lowering the rate of false positives.

Once feature extraction from a potentially malicious website is complete, an RDF model is generated from the HTML source code [8]. Twenty-one features were chosen where no equivalent websites exist to send the same element. There are now more vocabularies, such as Extensible Hypertext Markup Language (XHTML) [9], Hypertext Transfer Protocol (HTTP) [10], and Dublin Core [11, 12]. Random Forest is one of the most effective machine learning methods for categorization [13], where it maintains high accuracy even in the presence of less sensitive outliers and missing values in parameter choices. As can be seen in Figure 4, the system's performance was measured across three different parameters, with a True Positive Rate (TPR) of 98.8 percent, a False Positive Rate (FPR) of 1.5 percent, and an Accuracy (ACC) of 98.6 percent.

Fig 4: Comparison Result

The phishing detection method of visual resemblance is introduced by Ankit Kumar Jain and B. B. Gupta [14]. This research takes a comprehensive look at phishing attacks, their techniques, the best visual similarity approaches currently available for detecting phishing, and their survey-based methodology. The poll provided a more in-depth look at the problem, the current explanation space, and the framework for future study into how visual similarity-based tactics might be used to effectively deal with phishing incidents. To determine which Visual similarity-based strategies were more effective, a comparison was done. As can be shown in Table 1, the Hybrid strategy outperformed the other methods in terms of accurately detecting phishing assaults. However, as noted by the Authors, no single approach is sufficient when employing phishing.



Table 1: Comparative analysis of multiple types of hybrid attacks

Error on the part of Rishikesh Mahajan and Irfan Siddavatam. No matching citations were found. [15] Demonstrate how you would use machine learning to spot phishing attempts. This study investigates the performance of the machine learning methods Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) in detecting phishing URLs. The effectiveness of classifiers is measured by comparing data from the training set with data from the testing set. In Table 2, we can see that using 90% of the data as a training dataset, Random Forest was able to attain an accuracy of 97.14.



Table 2: Classifier’s performance

A machine-learning strategy for detecting phishing websites is presented by Kahksha and Sameena Naaz [13]. The focus of this study was on developing a model to safeguard users from phishing scams. In this research, a phishing dataset was analyzed using Decision Tree, Linear Model, Random Forest, and Neural Network techniques. MillerSmiles, PhishTank, and Google search operators were used to compile the dataset. There are a total of 2456 occurrences and 30 characteristics available in the dataset. Attribute values can be any integer between -1 and 1, where -1 indicates phishing, 0 indicates suspicion, and 1 indicates legitimacy. After the dataset has been cleaned and transformed into the desired shape, it is divided into two sections for training and testing, each comprising 70% of the total. In this, we used four different machine learning algorithms: the decision tree, the random forest, the neural network, and the linear model [16]. Data from the UCI Machine Learning Repository was used to conduct an analysis of 2456 websites, each of which was characterized by 30 attributes. Here, Rstudio has been utilized for the actual coding. On the aforementioned dataset, we tested the accuracy of the decision tree, the random forest, the neural network, and the linear model in terms of the true positive rate, the false negative rate, the F measure, and the rate of false positive. These figures are the result of a combination of the confusion matrix from the testing data and the performance indicator calculations. Fig 5 demonstrates that, over a variety of parameters, the Random Forest algorithm outperforms its rivals. Its 95.70% accuracy was the greatest of any algorithm tested, beating out that of DT (90.4%), NN (90.7%), and LM (92.1%).

Fig 5: Comparison of Existing Algorithms

Several machine learning for intrusion detection approaches [17] [18] [19] [20] have been presented, with a focus on logs from various sources. This exemplifies the power of machine learning techniques for anomaly detection and their utility in several domains of information security.

1. **DISCUSSION**

This section gives an overview of the study, including its scope and objectives, and explains the background material and specific methods that will be employed to achieve those goals. Several anti-phishing technique researchers contributed to this survey, presenting different algorithms and methods for categorizing phishing websites. Reading through the articles, it becomes clear that most of the work is accomplished with the use of machine-learning techniques. Therefore, the results of the most recent dataset analysis and classification using machine learning demonstrate the superiority of machine learning over competing methods.

The intelligent machine learning algorithm is a computational innovation that falls under the umbrella of the artificial intelligence subfield known as machine learning. This technique is effective in large datasets, where it mitigates the drawback of the new method and the risk of a zero-day assault. Classifiers based on machine learning were able to attain an accuracy of over 99%, demonstrating their superiority over other approaches. The quality of results varies depending on the size of the training data, the features used, and the classifier employed. Kahksha and Sameena Naaz's earlier research offered a diagram depicting the architecture of their proposed work (Fig 6).



Fig 6 Machine Learning Framework for the Proposed Work

1. **CONCLUSION**

In a nutshell, phishing detection systems are essential for user safety, preventing consumers from falling victim to online fraud, sharing sensitive information with an attacker, and successfully employing phishing as an attack method. Many of the currently available phishing detector tools make false positives because the blacklist isn't constantly being updated in real time without human interaction. The inaccuracy of detection is often to blame for such problems. Due to the importance of these issues, numerous researchers have explored various methods to enhance the reliability of phishing attack detection while reducing the frequency of false positives. Due to the randomness and ever-evolving URL patterns of attacks, the reference model will need to be changed at some point. The purpose of this review article is to examine and analyze existing literature in order to locate a workable method for categorizing phishing and genuine websites. The literature evaluation supports the conclusion that Machine Learning is the best method for accomplishing the paper's goals.

**REFERENCES**

1. R. Zieni, L. Massari and M. C. Calzarossa, "Phishing or Not Phishing? A Survey on the Detection of Phishing Websites," in IEEE Access, vol. 11, pp. 18499-18519, 2023.
2. Abbasi, A., Dobolyi, D. G., Vance, A., & Zahedi, F. M. (2021). The Phishing Funnel Model: A Design Artifact to Predict User Susceptibility to Phishing Websites. Information Systems Research.
3. Mahmoud Khonji, Y. I. (2013). Phishing Detection: A Literature Survey.
4. Bhawna Sharma, P. S. (2019). PhishAlert: An Efficient Phishing URL Detection via Hybrid Methodology. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8.
5. Gururaj Harinahalli Lokesh & Goutham BoreGowda (2020): Phishing website detection based on effective machine learning approach, Journal of Cyber Security Technology
6. Ushamary Sharma, B. S. (2015). Phishing-An Analysis on the Types, Causes, Preventive Measuresand Case Studies in the Current Situation. National Conference on Advances in Engineering, Technology & Management (AETM'15), Volume: IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,p-ISSN: 2278-8727, PP 01-08.
7. Vamsee Muppavarapu, A. R. (2018). Phishing Detection using RDF and Random Forests. The International Arab Journal of Information Technology, Vol. 15, No. 5.
8. Kremic E. and Subasi A.,(2015) “Performance of Random Forest and SVM in Face Recognition,” The International Arab Journal of Information Technology, vol. 13, no. 2, pp. 287-293.
9. Ali, W. (2017). Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection. International Journal of Advanced Computer Science and Applications, 8(9), 72-78.
10. Subasi, A., Molah, E., Almkallawi, F., & Chaudhery, T. J. (2017). Intelligent phishing website detection using random forest classifier. In Electrical and Computing Technologies and Applications (ICECTA), 2017 International Conference on (pp. 1-5). IEEE.
11. Samar Muslah Albladi, George R. S. Weir (2018)“User Characteristics that Influence Judgment of Social Engineering Attacks in Social Networks”, Human-centric Computing and Information Sciences, SpringerOpen, Vol. 8, No. 1, pp. 1-24.
12. 32. Tore Pedersen, Christian Johansen, Audun Josang, (2018)“Behavioural Computer Science: An Agenda for Combining Modelling of Human and System Behaviours”, Human-centric Computing and Information Sciences, SpringerOpen, Vol. 8, No. 1, pp. 1-20.
13. Abdelhamid N, Thabtah F, Abdel-jaber H(2017). Phishing detection: a recent intelligent machine learning comparison based on models content and features. In Beijing, China: IEEE; 2017
14. Ankit Kumar Jain, B. B. (2017). Phishing Detection: Analysis of Visual Similarity Based Approaches. In Springer Science+Business Media, LLC, part of Springer Nature.
15. Ankit Kumar Jain, B. B. (2016). A novel approach to protect against phishing attacks at client side using auto-updated whitelist. EURASIP Journal on Information Security, Springer Open,, Vol. 2016, No. 1, pp. 1-11.
16. Wenbin Yao, Yuanhao Ding, Xiaoyong Li.(2018).“Deep Learning for Phishing Detection”, 2018 IEEE International Conference on Parallel & Distributed Processing with Applications, Ubiquitous Computing & Communications, Big Data & Cloud Computing, Social Computing & Networking,
17. Asif-Iqbal, H., Udzir, N. I., Mahmod, R., & Ghani, A. A. A. (2011). Filtering events using clustering in heterogeneous security logs. Information Technology Journal, 10(4), 798-806.
18. Hajamydeen, A. I., Udzir, N. I., Mahmod, R., & GHANI, A. A. A. (2016). An unsupervised heterogeneous log-based framework for anomaly detection. Turkish Journal of Electrical Engineering & Computer Sciences, 24(3), 1117-1134.
19. Hajamydeen, A. I., & Udzir, N. I. (2016). A refined filter for UHAD to improve anomaly detection. Security and Communication Networks, 9(14), 2434-2447.
20. Hajamydeen, A. I., & Udzir, N. I. (2019). A detailed description on unsupervised heterogeneous anomaly based intrusion detection framework. Scalable Computing: Practice and Experience, 20(1), 113-160.