ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol19, Issue 1,2025

AI-POWERED PHISHING WEBSITE DETECTION METHOD

[1]ALONE ANUSHA, [2]MARRIPELLY ARAVIND,[3] MUSKU ANVITHA, [4]MAMIDI MANISH KUMAR, [5]P MOUNESH, [6]Mr. DANDU SRINIVAS, [7]Mrs. DUNGA SIMHANA,

[12345]Student Department of DS, Narsimha Reddy Engineering College, Maisammaguda (V), Kompally, Secunderabad, Telangana-500100.

[6]Assistant Professor, Department of CSE, Narsimha Reddy Engineering College, Maisammaguda (V), Kompally, Secunderabad, Telangana-500100.

[7]Assistant Professor, Department of Mechanical Engineering, Narsimha Reddy Engineering College, Maisammaguda (V), Kompally, Secunderabad, Telangana-500100.

Abstract-

Nowadays, the Internet is an essential tool for both our personal and professional lives. Because of this, more and more people are opting to shop online. This reality leaves internet users open to a myriad of cyber hazards. Financial loss, credit card theft, data breaches, brand reputation issues, and consumer skepticism towards online banking and shopping are all possible outcomes of these dangers. One kind of cyber hazard is phishing, which is when a criminal creates a fake website in order to trick users into giving up personal information (such as passwords, usemames, and credit card numbers). At the heart of this study are methods for identifying phishing attempts. The researchers in this study used a machine learning method to identify phishing attempts. Consequently, phishing may be accurately detected in this research.

Key words: phishing, attack; phishing; website detection; malware; machine learning.

INTRODUCTION

One common kind of cyber hazard is the phishing assault, which uses any kind of communication channel to deceive people into giving up personal information. In order to steal information that might do them or their companies harm, attackers use deception and make victims fall into their traps. The choice of communication channel is determined by the attacker's purpose and the kind of data. the first Threats of account deletion and demands for ransom are also part of it. Customer information such as passwords and credit card details may be compromised using another deceitful tactic known as email spoofing. The primary goal of phishing is to steal sensitive information, such as login passwords for online banking or credit card numbers. Online firms' reputations take a hit as a result of these fraudulent operations, which weaken confidence in online transactions. Computer systems are still susceptible to assaults, even with data encryption techniques. [2] in Avoiding phishing attacks requires awareness and attentiveness. To avoid danger, make it a habit to carefully browse the web and check the legitimacy of connections. Software and add-ons for web browsers may detect and block malicious websites that try to steal login information. Security is improved by implementing network systems that restrict access to only authorized sites; nevertheless, this approach compromises user comfort. [1] To detect phishing attempts, this study used machine learning. The methods used to detect phishing websites are heuristic-based and gather data from websites in order to determine their legitimacy. Heuristics, in contrast to blacklists, can identify phishing sites while they are being built in real time. When it comes to distinguishing between various kindsof websites, effective heuristic approaches depend on discriminating criteria. Phishing websites may be identified using the heuristic technique by analyzing HTML or URL signatures. The efficacy of this approach is being investigated in ongoing research. the third Machine learning and data mining methods are evaluated for their ability to anticipate phishing sites. Among these algorithms are Logistic Regression (LR), Bayesian Additive Regression Trees (BART), Classification and Regression Trees (CART), Random Forests (RF), and Neural Networks (NN). To train and test classifiers, experiments were

ISSN 2454-9940

www.ijasem.org

Vol19, Issue 1,2025

conducted using a dataset consisting of 1,172 phishing emails and 1,718 legitimate emails, using 43 different functions. From the results, we can see that RF had the best accuracy rate at 7.72%, followed by CART at 8.13%, LR at 8.58%, BART at 9.69%, Support Vector Machines (SVM) at 9.90%, and NN at 10.73%. But the results show that no one classifier is better than the others in identifying phishing sites. [4] Bagging, AdaBoost, SVM, CART, NN, RF, LR, NB, and BART are some of the machine learning-based detection techniques (MLBDMs) that are examined and compared in this study. There are 1,500 legitimate websites and 1,500 malicious ones in the sample. A total of eight elements make up CANTINA's evaluation factors. [4]

Phishing Website Attacks and Trends

During the reporting month of December 2021, the Anti-Phishing Working Group (APWG) recorded 316,747 assaults. At the start of the year 2020, phishing schemes became more common. There was a 6.5% increase in the frequency of phishing scams targeting bitcoin exchanges and wallet providers in the fourth quarter of 2018, with the banking sector being the most targeted. The number of businesses found to have fallen victim to ransomware increased by 36% between the third and fourth quarters. When business users looked into phishing emails, they found that 51.8% were trying to steal credentials, 38.6% were response-based attacks (including BEC, 419, and gift card scams), and 9.6% were trying to deliver malware.

Also, in December of 2021, there were 316,747 assaults recorded by the APWG. Throughout the APWG's reporting history, this is the highest monthly total. Since the beginning of 2020, phishing schemes have become increasingly common. With 23.2% of all assaults occurring in the fourth quarter of 2018, the financial sector was the most often attacked by plrislring. There has been a consistent lack of cyberattacks targeting webmail and software as a service providers. The proportion of assaults consisting of plrislring scams targeting bitcoin exchanges and wallet providers increased to 6.5%. From Q3 to Q4, there was a 36% increase in the number of companies found to have been infected by ransomware. Of the emails reported by business users, 51.8% were credential theft plrislring attacks, 38.6% were reaction based attacks (such as Business email composite (BEC), 419, and gift card scams), and 9.6% were elsewhere.

Website Security Flaws The most common Internet security flaws that lead to spear phishing attacks are covered in this section: Vulnerabilities in Cross-Site Scripting (XSS) happen when malicious scripts are embedded into web pages that users see. This might cause the victim's browser to execute unauthorized code. As a result of tlris, hackers may obtain login credentials or divert users to phishing websites. D. Stuttard and M. Pinto's "The Web Application Hacker's Handbook" is an excellent resource for learning about and fixing cross-site scripting (XSS) vulnerabilities. [6]. Next, attackers may take advantage of Cross-Site Request Forgery (CSRF) vulnerabilities to sneakily execute operations on a certain website without the user's knowledge or permission. Subtitling fonts or making purchases on plrislring sites are examples of plris. A. Barth et al.'s "Robust Defenses for Cross-Site Request Forgery" offers helpful information on how to protect yourself against CSRF attacks. [7] Additionally, SQL injection vulnerabilities develop when malicious actors are able to manipulate user-supplied data in order to perform SQL queries against a website's database. Prism attacks may take advantage of these vulnerabilities, which can provide hackers access to sensitive user data. [8] When hostile actors get the session ID of a user without authorization, they may impersonate the user and do harmful acts, such as diverting them to fraudulent websites, which is known as session hijacking. The techniques and effective countermeasures for session hijacking are covered in the article "Session Hijacking and Its Countenneasures" [9] by M. Naveed et al. 236 total III.

PROCEDURE

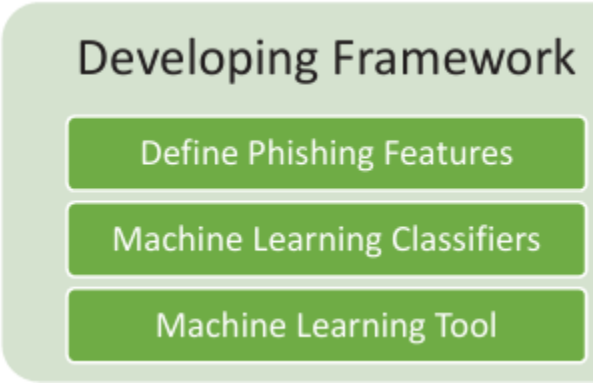
Because it is possible to roll back to earlier phases with little loss and apply new research advances, the tlris research technique is being applied in this work. Not only that, but if issues emerge at this level, the method allows for modifications to any phase to address them. Finally, researchers may easily adapt this research technique to meet the needs of the study subject. Hey there!

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN 2454-9940

www.ijasem.org

Vol19, Issue 1,2025



The Characteristics of Phishing Attacks Characteristics determined by URLs will be the main focus. The URL is the first thing to check while deciding whether or not to plrish a website. Domain URLs that are unique exhibit certain characteristics. In order to get traits linked to these locations, the URL is examined. As part of this research, we will be studying the following URL-based features: 1. Take Care of Bar-Based Features n. Base Features of Abnonnal HTML and JavaScript-based Functionality tv. Functions Exclusive to a Domain

Machine Learning Classifiers and Tools

Machine learning, a subfield of artificial intelligence, is able to improve existing systems and anticipate future events without human intervention. Classifiers, which are widely utilized in intrusion detection systems, affect both the learning process and the outcomes of predictions. There are two main methods for machine learning: supervised and unsupervised. In order to minimize mistakes, this study utilizes supervised machine learning with tagged data (both normal and phishing). We compare five classifiers: RF, J48, Naive Bayes, Logistics, and K-Nearest Neighbors (KNN). When it comes to classification or regression, Random Forest is a great collective learning approach since it trains numerous decision trees.

Data set training makes advantage of Google Colab's flexibility and cloud capabilities. Python machine learning makes good use of it. Crucial to the memory-hogging machine learning algorithm's optimization is the distribution of GPU assets from Google servers to otherwise restricted hardware on the programmer end. The data set is saved in Google Storage, which is a cloud drive architecture. Then, it is imported into the Colab online notebook and trained. After training the model, it is loaded into the Pi and tested using the data that has been acquired. section 4 [15] [16]

DYNAMIC ANALYSIS OF PHISHING WEBSITE AND DETECTION TECHNIQUE

The four parts that comprise the design model are data collection, factor identification, model testing, and result comparison. In the subject, each component will get a brief review.

Figure 2. A Method for Identifying Phishing Websites using Dynamic Analysis Here, the design model will be used to put the proposed solution into action. At this stage, the first order of business is to set up a desktop computer, laptop, or mobile device with the study-specific software, such Google Colab.

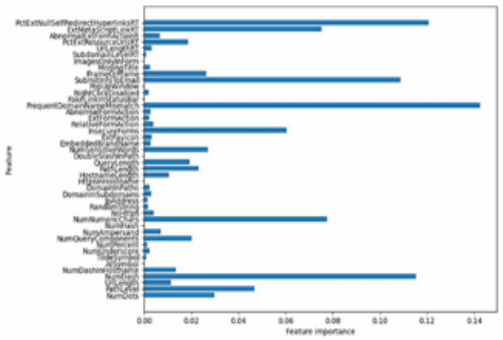
The four-step empirical evaluation process is at the heart of the dynamic analysis method: You can see the dynamic analysis process for phishing website detection in Fig. 2. Gathering datasets of phishing websites is the first stage in identifying such sites. In their studies and trials, researchers often use datasets from numerous phishing websites. The PhishTank dataset is one example of a popular tool for reporting and validating phishing URLs[l7]. A publicly accessible dataset including a comprehensive collection of phishing URLs is provided by OpenPhish [l8]. A database of reported phishing URLs is kept by the Anti-Phishing Working Group (APWG)[l9]. To facilitate benchmarking and model building, platforms such as Kaggle house

ISSN 2454-9940

www.ijasem.org

Vol19, Issue 1,2025

community-contributed datasets of phishing websites[20]. In order to study phishing patterns and improve detection methods, researchers might look into the publicly accessible datasets hosted on GitHub [21]. Researchers may use these databases to study phishing trends, create better detection algorithms, and test how well their methods work. The four-step empirical evaluation process is at the heart of the dynamic analysis method: You can see the dynamic analysis process for phishing website detection in Fig. 2. As a first step in identifying phishing websites, gathering datasets of such sites is essential. New research shows that the accuracy of an experiment improves as the number of datasets utilized grows. references [22], [23] being Kaggle and similar phishing datasets are used often by researchers.



Listing of Features (Fig. 3) The components of the malicious website were then classified according to the essential characteristics they had. This study has utilized the feature selection approach to identify the important parameters for reliable phishing website detection. To make sure that a pattern stands out between legitimate websites and phishing ones, many tools are used. In Figure 3, you can see the study team's inventory of the phishing website traits they looked at.

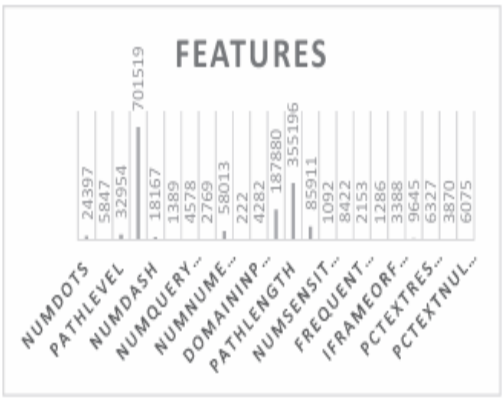


Fig. 4. Chosen Features The following step is defining phishing features by extracting the most frequent data used from the dataset in Fig. 3 using correlation attribute evaluation which involves

appraising a trait's worth by looking at how closely it ties to the category. Besides providing a rating from best to worst, it also displays the rank number for every quality [l2]. Figure 4 shows that certain attributes are more highly ranked than others due to their frequency of application in the detecting process. After all the parts are in place, the third stage is to test the dataset that will be used to evaluate the experiment. Testing and evaluation are conducted to solve the issue statement and see whether the restriction of existing journals is avoided. The major goal of this analysis is to prove the efficacy of the proposed detection model so that the results and claims of this study can be confirmed. Also, by evaluating and testing, the research experiment might find limitations and problems, which allows for more tweaks to get the right result. The last thing to do is to examine the outcomes using machine learning methods like logistic regression[30], Naive Bayes[28], KNN[29], and random forest[26, 27]. These techniques mine the data for hidden meaning and use it to inform their predictions. Accurate outcomes are generated by random forest via the construction of many decision trees. J48 is a decision tree classifier that finds important traits and gives rules that may be understood. One effective probabilistic method for big datasets is Naive Bayes. KNN uses the distance between new occurrences and existing ones to make classifications. To classify data into two or more categories, logistic regression models the connections between the variables. These methods make it easier to find patterns, get insights, and make well-informed decisions.

Discussion

The findings show the output of the following machine learning classifiers: Logistic, Naive Bayes, KNN, and Random Forest. Furthermore, the accuracy, precision, and recall metrics implemented in Python were used in this examination of the different measurements. The results from the testing set, which included 25 characteristics of phishing websites and five selected classifiers, are shown in Table I.

Table I. The Related Study Classifiers Analysis

Classifiers	Accuracy	Precision	Recall	FPR	TP
Random Forest	94.10%	0.978	0.904	0.021	0.9
J48	92.10%	0.917	0.926	0.084	0.9
Naïve Bayes	83.00%	0.921	0.771	0.071	0.7
KNN	92.21%	0.923	0.921	0.079	0.9
Logistic	89.50%	0.895	0.895	0.105	0.8

eXIstmg tactics need more refinement. Chapter 3 presents a recommended technique that tries to help internet users identify phishing websites. This chapter also gives a detailed rundown of the study process, including all the methods and equipment that were used. To guarantee the efficacy of the approach for detecting phishing websites, the next chapter will discuss the procedures for installation, testing, and assessment. Chapter 4's results also show that the Random Forest algorithm beat the competition with a stunning accuracy rate of over 94%, as well as with precision, True Positive Rate (TPR), and Receiver Operating Characteristic (ROC) values. Similarly, J48 and KNN algorithms consistently performed over 90%, although Nai:ve Bayes and Logistics performed somewhat worse in several tests. These findings suggest that the Random Forest algorithm performs the best when it comes to identifying phishing attempts. With the internet's transformative power on people's lives, the research stresses the need of dealing with security concerns like phishing. Using a random forest classifier and optimizing feature datasets, this study achieves excellent accuracy in detecting phishing websites that are based on machine learning. Feature selection, lowering the false alarm rate, and investigating dynamic analysis methods are some of the improvement topics highlighted in the research. Improving the detection methods and prioritizing important feature selection should be the goals of future study, with dynamic analysis techniques also being considered.

REFERENCES

[1]. S. Hossain, D. Sarma, and R. J. Chakma, "Machine Learning-Based Phishing Attack Detection," 2020. [Online]. Available: www.ijacsa.thesai.org

[2]. M. N. Alam, D. Sarma, F. F. Lima, I. Saha, R. E. Ulfath, and S. [3] [4] [5] Hossain, "Phishing attacks detection using machine learning approach," in Proceedings of the 3rd International Conference on Smart Systems and Inventive Technology, ICSSIT 2020, Institute of Electrical and Electronics Engineers Inc., Aug. 2020, pp. 1173 1179. doi: 10.1109/ICSSIT48917.2020.9214225.

[3]. N. Chou, R. Ledesma, Y. Teraguchi, and J. C. Mitchell, "Client-side defense against web-based identity theft." [Online]. Available: www.ebaymode.com

[4]. D. Miyamoto, H. Hazeyama, andY. Kadobayashi, "An Evaluation of Machine Learning-based Methods for Detection ofPhishing Sites." "Phishing E-mail Reports and Phishing Site Trends 4 Brand-Domain Pairs Measurement 5 Brands & Legitimate Entities Hijacked by E mail Phishing Attacks 6 Use of Domain Names for Phishing 7-9 Phishing and Identity Theft in Brazil 10-11 Most Targeted Industry Sectors 12 APWG Phishing Trends Report Contributors 13 4 th PHISHING ACTIVITY TRENDS REPORT," 2022. [Online].

[5]. W. G. J. Halfond, J. Viegas, and A. Orso, "A Classification of SQL Injection Attacks and Countermeasures," 2006.

[6]. L. and M. A. Vishnoi, "International Journal of Computer Science & Information Security," International Journal of Computer Science & Information Security, vol. 15, pp. 1-425, 2013, [Online].Available: https :/ /sites.google.com/site/ijcsis/

[7]. "Research Methodology Methods and Techniques (PDFDrive)".

[8]. Choon Lin Tan, "Phishing Dataset for Machine Learning: Feature Evaluation," M endeley Data, Mar. 24, 2018.

[9]. Betha Nurina Sari, "CorrelationAttributeEval," ResearchGate, Apr. 25, 2017.

[10]. F. Vanhoenshoven, G. Napoles, R. Falcon, K. Vanhoof, and M. Koppen, "Detecting malicious URLs using machine learning techniques," in 2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016, Institute of Electrical and Electronics Engineers Inc., Feb. 2017. doi: 10.11 09/SSCI.2016. 7850079.

[11]. Institute of Electrical and Electronics Engineers, IEEE Communications Society;, Denshi Joho Tsiishin Gakkai (Japan). Tsiishin Sosaieti, and Han'guk T'ongsin Hakhoe, ICUFN 2019 : the lith International Conference on Ubiquitous and Future Networks : July 2 (Tue.)-July 5 (Fri.) 2019, Zagreb, Croatia.

[12]. M. Kuroki, "Using Python and Google Colab to teach uundergraduate microeconomic theory," International Review of Economics Education, vol. 38, p. 100225, Nov. 2021, doi: 10.10 16/J.IREE.2021.1 00225. [16] Prabanjan Raja, "What is Google Colab?," Scaler Topics, Feb. 11, 2022