International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 05 | May - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

PHISHING WEBSITE DETECTION

Vijayalakshmi.J[1],Aathikesavan.A[2],Deepika.S[3],Kowsalya.P[4]

1*Assistant Professor, Department Of Artificial Intelligence and Data Science,
Sri Shakthi Institute of Engineering and Technology, Coimbatore.

2,3,4.Department of AI&DS,
Sri Shakthi Institute of Engineering and Technology, Coimbatore.

ABSTRACT

Phishing is an internet scam in which an attacker sends out fake messages that look to come from a trusted source. A URL or file will be included in the mail, which when clicked will steal personal information or infect a computer with a virus.The goal was to get as many people to click on a link or open an infected file as possible. There are various approaches to detect this type of attack.One of the approaches is machine learning. The URL's received by the user will be given input to the machine learning model then the algorithm will process the input and display the output whether it is phishing or legitimate.

Keyword: Phishing, Website Detection, Cybersecurity, Fraud, URL Analysis, SSL Certificate, Behavioral Analysis, Machine Learning, Cyber Threats, Malicious Websites, Security Measures

INTRODUCTION:

Phishing has emerged as the biggest issue, affecting people, businesses, and even entire nations. The availability of several services, including social networking, software downloads, online banking, entertainment, and education, has sped up the development of the Web recently. Consequently, a vast quantity of data is continuously downloaded and uploaded to the Internet. Social engineering techniques use spoof emails that look like they are from reliable companies and agencies to send visitors to phony websites that trick them into providing financial information such usernames and passwords. Technical techniques include installing malicious software on computers in order to directly steal credentials; these systems are commonly used to intercept users' passwords and usernames for online accounts.

The widespread adoption of the internet in many facets of daily life has yielded many advantages, but it has also created new cybersecurity challenges. Of the various threats, phishing attacks are particularly noteworthy as a sophisticated and widespread method of taking advantage of people and organizations. Phishing is the practice of deceiving users into divulging sensitive information, such as passwords, financial information, or personal data, by means of deceptive techniques. As phishing tactics become more sophisticated, there is an increasing demand for sophisticated and adaptive mechanisms to detect and stop these malicious activities.

Static rules and signature-based techniques have been widely used in traditional phishing detection methodologies. Nevertheless, these approaches may find it difficult to stay up with the constantly changing tactics that cybercriminals use. A potential solution to this problem is to use machine learning (ML) approaches, which use algorithms to examine dynamic data and trends connected to phishing websites.

This research presents a novel machine learning-based method for phishing website identification. Our suggested methodology seeks to improve the precision and versatility of phishing detection by fusing sophisticated feature extraction techniques with a wide range of machine learning algorithms. Because online threats are always changing, a proactive defense system is necessary. To keep the model up to date with new phishing techniques, continuous learning methods have

been incorporated

LITERATURE REVIEW:

Phishing detection has garnered a great deal of research interest, particularly in view of the ever-complex threat landscape. As phishing attempts continue to evolve, researchers are depending more and more on machine learning. The dynamic nature of the threat landscape has led to a significant rise in research effort in the field of machine learning-based phishing detection, necessitating innovative and adaptable approaches. This review of the literature provides an overview of prior studies and approaches in the topic of phishing detection, with a focus on the use of machine learning techniques. By examining the contributions and limitations of prior research, this review aims to contextualize the current state of knowledge and identify knowledge gaps, laying the foundation for the recommended approach.

PROPOSED METHODOLOGY

A phishing website is a kind of social engineering where malicious websites and unified resource locators (URLs) are mimicked. The most typical method by which phishing attacks take place is through the Uniform Resource Locator (URL). The sub-domains of the URL are entirely under the phisher's control. Because the URL includes directories and file components, the phisher can change it.The linear-sequential approach, sometimes referred to as the waterfall model, was employed in this study. The waterfall technique is regarded as standard, yet it functions finest when there aren't many requirements. The program was broken up into smaller parts, each of which was constructed with hand-written code and frameworks.

TYPES

1.Deceptive Phishing:

This is the most common kind of phishing attack, where a cybercriminal poses as a reputable company, institution, or website in order to obtain sensitive personal data from the victim, including passwords, bank account details, credit card details, and login credentials. This type of attack is less sophisticated because there is no tailoring or customization for the individuals.

2.Spear Phishing:

This type of phishing email contains harmful URLs together with a lot of personalized data about the potential victim. The email may contain the name, position, company name, friends, coworkers, and other social media details of the recipient.

3.Whale Phishing:

This type of phishing targets corporate executives like CEOs and top-level management staff in an attempt to spear phish a "whale," in this case a top-level executive like a CEO.

4.URL Phishing:

The fraudster or cybercriminal uses a URL link to infect the target. Humans are gregarious beings; they will gladly accept friend requests by clicking the link, and they might even divulge private information like email addresses.

This is a result of users being redirected to a phony web server by phishers. Attackers also employ secure browser connections to execute their illegal activities. Businesses are unable to train their employees in this area because they lack the necessary tools to counter phishing assaults, which leads to a rise in phishing attacks.

As general defenses, businesses are encrypting critical data, updating all of their systems with the most recent security patches, and training employees through simulated phishing attacks. One of the easiest ways to fall prey to this phishing attack is to browse carelessly. Phishing websites have an identical design to legitimate websites.

DATA COLLECTION

The open-source program Phish Tank was used to compile the list of phishing URLs. This website offers an hourly updated collection of phishing URLs in multiple formats, such as csv, json, and others. Using 5000 randomly selected phishing URLs, this dataset is used to train machine learning models.

DATA CLEANING

To clean up the data, add missing values, smooth out crooked data, find and remove outliers, and fix anomalies.

DATA PREPROCESSING

The process of cleaning unstructured raw data to create a tidy, well-organized dataset that may be utilized for additional study is known as data preparation. Data

preprocessing is a cleaning procedure that turns unstructured raw data into a tidy, well-organized dataset that may be utilized in subsequent studies.

Before the ML model is trained, the data is divided into 8000 training samples and 2000 testing samples. The dataset clearly shows that this is a challenge involving supervised machine learning. The two primary categories of problems in supervised machine learning are classification and regression. This data collection has a categorization issue because the input URL is categorized as either phishing or legitimate. For the dataset training for this research, the following supervised machine learning models were investigated: Multilayer Decision Tree Perceptron, Random Forest, Autoencoder Neural Network , XGBoost and Support Vector Machines.

DATA SET

The datasets were gathered from Phishing Tank, an open-source platform. The gathered dataset was saved as a CSV file. The dataset consists of eighteen columns, which we changed using a data pre-processing technique. We familiarized ourselves with a few data frame approaches to see the features in the data. A few charts and graphs are provided for visualization, so that you can understand how the data is distributed and how different features relate to one another. A machine learning model's training is unaffected by the Domain column. 16 features and a target column are available today. There is no shuffling involved; the recovered features of the datasets from phishing and valid URLs are only concatenated in the feature extraction fileTo even out the distribution and divide the data into training and testing sets, we must shuffle the data. Additionally, overfitting during model training is eliminated.

ADDRESS BASED CHECKING

Below are the categories been extracted from address based

1.Domain of the URL : where the extracted domain that appears in the URL

2.IP Address in the URL :

It is verified if an IP address is present in the URL. An IP address may appear in URLs in place of a domain name. We can be positive that a URL is being used to gather sensitive data if an IP address is included in place of a domain name.

4."@" Symbol in URL :

It is verified if the URL contains the "@" symbol. The real address is typically discovered after the "@" sign in a URL since anything preceding the "@" symbol is ignored by browsers.

4.Length of URL :

determines the length of the URL. By utilizing a long URL, phishers can mask the dubious part of a URL in the address bar. In the context of this research, a URL is considered phishing if its length exceeds 54 characters..

5.Depth of URL :

determines the depth of the URL. This feature counts the number of subpages in the provided address based on the '/'.

6.Redirection "//" in URL :

It is verified if"//" is included in the URL. The URL route's use of the character "//" denotes a website redirection for the user. The URL's "//" location is determined through calculation. We found that the "//" should be in the sixth position if the URL starts with "HTTP." However, if the URL employs "HTTPS," the "//" ought to appear in the seventh position.

7.Http/Https in Domain name :

It is verified if "http/https" is present in the URL's domain portion. Phishers may add the "HTTPS" token to a URL's domain section in order to trick consumers.

8.Using URL Shortening Services :

A technique called URL shortening allows a URL to be shorter while still pointing to the target destination on the "World Wide Web." Using a "HTTP Redirect" on a short domain name that refers to a webpage with a lengthy URL is how this is done.

9.Prefix or Suffix "-" in Domain :

Verifying if the domain portion of the URL contains a '-'. The dash symbol is rarely used in real URLs. In order to create the appearance that they are interacting with a trustworthy website, phishers usually append prefixes or suffixes to domain names, separated by a semicolon (-).

HTML AND JAVA SCRIPT BASED CHECKING

It is possible to remove several of the elements that make up this group. Out of all of them, the following were taken into consideration for this project.

1.IFrame Redirection :

With the help of the HTML tag IFrame, you can add a different webpage to the one you are currently viewing.

Phishers can utilize the "iframe" tag to remove the frame's borders and make it invisible. In this instance, phishers use the "frame border" element, which prompts the browser to draw a visual border.

2.Status Bar Customization :

Phishers may utilize JavaScript to trick visitors into seeing a false URL in the status bar. To get this feature, we'll need to delve into the webpage source code, specifically the "on Mouseover" event, and see if it alters the status bar.

3.Disabling Right Click :

Phishers employ JavaScript to block the right-click feature, preventing users from seeing and saving the source code of the webpage. The handling of this functionality is similar to that of "Hiding the Link with on Mouseover." However, we'll search the webpage source code for the {{event"event. button==2" and check if right-click functionality is disabled.

4.Website Forwarding :

Phishing and legitimate websites can be distinguished from one another by counting the number of times a website has been redirected. We found that in our sample, genuine websites were routed just once. On the other hand, phishing websites that use this feature have at least four redirects.

5.Implementation :

In this section of the report, we will look at our artefact's implementation component, with a particular emphasis on the generated solution's description. For this kind of work, supervised machine learning is necessary.

6.List-Based Approaches:

In 2016, Jain and Gupta presented a whitelist-based, automatically updated method to defend against client-side phishing attempts. Its 86.02% accuracy and less than 1.48% false-positive rate, which suggests a false alert for phishing assaults, are shown by the trial findings. This method's quick access time ensures a real-time environment and products, which is another advantage.

7.Heuristic Strategies:

Three phases make up the PhishWHO phishing detection technique, which was introduced by Tan et al. It first gathers identity keywords from the page's HTML using a weighted URL token mechanism, then groups the N-gram model. Second, it uses the keywords to locate the official website and the legal domain in popular search engines. The target website's domain is then compared to the legitimate domain to see if it is a phishing website or not. To determine whether the website was legitimate, Chiew et al. used a picture of the logo from the website. In this study, the authors used machine learning techniques to extract a logo from web page photos. They then used the logo as a keyword to query the domain using the Google search engine. Consequently, this category was also referred to as search engine-based strategy by certain studies.

8.Machine Learning-Based Methods:

In comparison to other approaches, machine learning-based countermeasures are suggested as a more accurate way to combat dynamic phishing assaults with a reduced rate of false positives. As a result, the six parts of the machine learning approach are feature extraction, data extraction, model training, model testing, and prediction.each part's flowchart. This flowchart serves as the foundation for machine learning-based phishing website detection solutions now in use, which optimize one or more components for improved performance.

9.Tiny URL Detection:

Rule-based feature selection strategies may not work for short URLs since they do not give the genuine domain, resource direction, or search parameters. It is challenging to convert tiny URLs produced by various services back to their original URLs. Moreover, character-level information cannot be extracted from tiny URLs by natural language processing because they are short strings.. Tiny URLs have the potential to trigger false or missed alarms if they are not specifically handled during data cleansing and preprocessing. The user experience of internet products is also crucial, and consumers are sensitive to false warnings from Internet security solutions.Real-Time System Response Time Rule-based models rely on third-party services from a URL string and rule parsing. As such, in a real-time prediction system that takes a single URL string as an input in every client request, they require a comparatively long response time. Phishing attacks have expanded to encompass a range of communication channels and target devices, including smart devices and personal PCs. For developers, covering all devices with a single solution is a significant task. To lessen the complexity of system development and the associated maintenance expenses,

language independence and operating environment independence should be taken into account.

MACHINE LEARNING MODELS

1. DECISION TREE CLASSIFIER

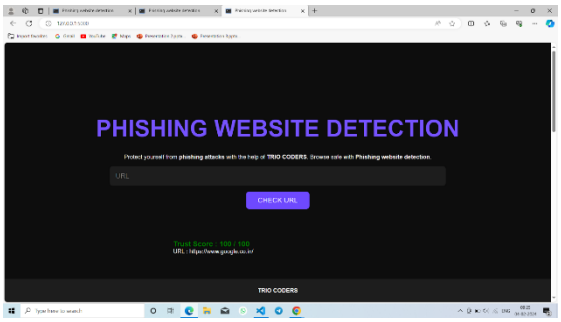
Decision trees are popular models for applications involving regression and classification. In essence, they are taught a hierarchy of if/else questions that result in a selection. The process of learning a decision tree involves committing the series of if/else questions to memory that, when answered correctly, will take the least amount of time. In order to construct a tree, the algorithm iterates over all possible tests and selects the most informative one about the target variable.

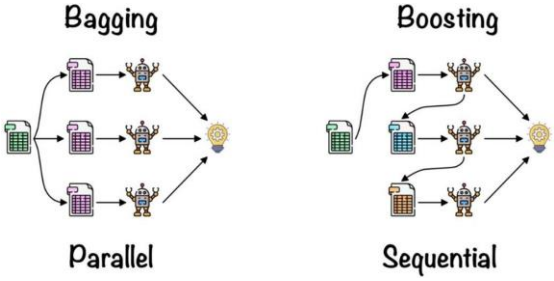
1. RANDOM FOREST

One of the most popular machine learning techniques for regression and classification is random forests. All that is a random forest is an arrangement of decision trees, each varying slightly from the others. The idea behind random forests is that, even if each tree might forecast rather well, on some data it will almost definitely overfit. They often function well with few parameter adjustments, are quite powerful, and don't require scalable data.

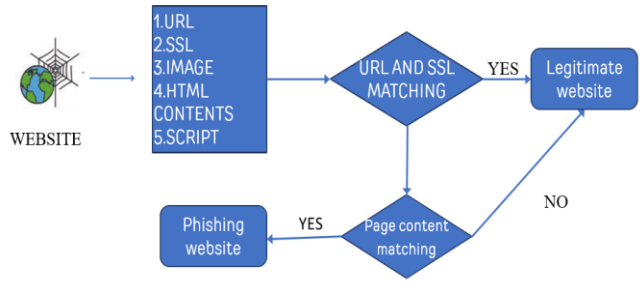
1. Bagging: Using replacement, it generates a distinct training subset from sample training data, and majority vote determines the final result. Take Random Forest, for instance.

2. Boosting: By building successive models with the maximum accuracy possible, this method turns ak learners become strong learners. As an illustration,ADA BOOST, XG BOOST.



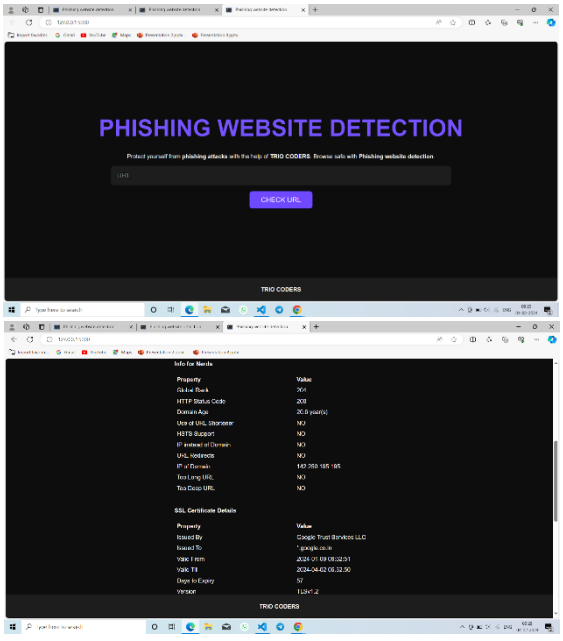


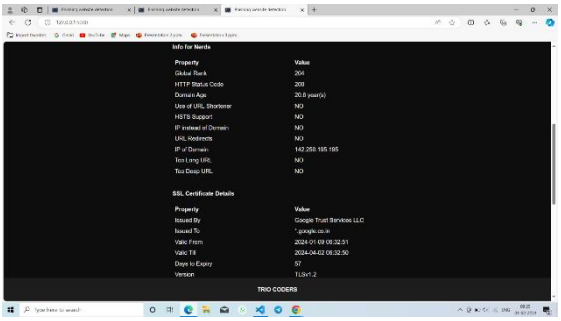
FLOW DIAGRAM



OUTPUT

Thus using the machine learning algorithms and following up the methodologies as said, we have successfully produced a system where we can accurately detect the Phishing website.





CONCLUSION

This survey featured a variety of methods and algorithms developed by multiple machine learning researchers to identify phishing websites.After going over the articles, we concluded that the majority of the work was completed with well-known machine learning methods, such as Random Forest, SVM, Naïve Bayesian, and Decision Tree.A new detection method similar to Phish Score and Phish Checker was suggested by some authors. Features pertaining to recall, accuracy, precision, and other aspects were combined. Successful experimental methods for identifying the URLs of phishing websites were compiled. Phishing websites are growing daily, thus in order to identify them, some elements may be added or removed.

REFERENCES

[1] 'APWG | Unifying The Global Response To Cybercrime' (n.d.) available: https://apwg.org/

[2] 14 Types of Phishing Attacks That IT Administrators Should Watch For [online](2021)https://www.blog.syscloud.com,available:https://www.blog.syscloud.comtypes-o f-phishing/

[3] Lakshmanarao, A., Rao, P.S.P., Krishna, M.M.B. (2021) 'Phishing website detection using novel machine learning fusion approach', in 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), Presented at the 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), 1164–1169

[4] H. Chapla, R. Kotak and M. Joiser, "A Machine Learning Approach for URL Based Web Phishing Using Fuzzy Logic as Classifier", 2019 International Conference on Communication and Electronics Systems (ICCES), pp. 383-388, 2019, July

[5] Vaishnavi, D., Suwetha, S., Jinila, Y.B., Subhashini, R., Shyry, S.P. (2021) 'A Comparative Analysis of Machine Learning Algorithms on Malicious URL Prediction', in 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Presented at the 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 1398–1402

[6] Microsoft, Microsoft Consumer safety report. https://news.microsoft.com/en-sg/2014/02/11/microsoft-consumersafety-index-reveals- impact-of-poor-online-safety-behaviours-in-singapore/sm.001xdu50tlxsej410r11kqvks u4nz.

[7] Internal Revenue Service, IRS E-mail Schemes. Available at https://www.irs.gov/uac/newsroom/consumers-warnedof-new-surge-in-irs-email-schem es-during-2016-tax-season-tax-industry-also-targeted.

[8] Abu-Nimeh, S., Nappa, D., Wang, X., Nair, S. (2007), A comparison of machine learning techniques for phishing detection.
Proceedings of the Anti-phishing Working Groups 2nd Annual ECrime Researchers Summit on - ECrime '07. doi:10.1145/1299015.1299021.

[9] E., B., K., T. (2015)., Phishing URL Detection: A Machine Learning and Web Mining-based Approach. International Journal of Computer Applications,123(13), 46-50. doi:10.5120/ijca2015905665.

[10] Wang Wei-Hong, L V Yin-Jun, CHEN Hui-Bing, FANG Zhao-Lin., A Static Malicious Javascript Detection Using SVM, In Proceedings of the 2nd International Conference on Computer Science and Electrical Engineering (ICCSEE 2013).

[11] Ningxia Zhang, Yongqing Yuan, Phishing Detection Using Neural Net-work, In Proceedings of International Conference on Neural Information Processing, pp. 714–719. Springer, Heidelberg (2004).

[12] Ram Basnet, Srinivas Mukkamala et al, Detection of Phishing Attacks: A Machine Learning Approach, In Proceedings of the International World Wide Web Conference (WWW), 2003.

[13] Sci-kit learn, SVM library. http://scikit-learn.org/stable/modules/svm.html.

[14] Sklearn, ANN library. http://scikit-learn.org/stable/modules/ann.html.

[15] Sclera,Randomforestlibrary.http://scikitlearn.org/stable/modules/Randomforesets.h tml.