WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Phishing Website Detection

Prof. Vaishnavi Ganesh[1], Kaustubh Shastrakar[2], Anjali Sulakhiya[3], Shrutika Satange[4], Jay Dhurat[5], Nikhil Miralwar[6]

[1,2,3,4,5,6]Department of Computer Science and Engineering, Priyadarshini College of Engineering, Nagpur, India

ABSTRACT –

Phishing websites are a major threat to online security. These fraudulent websites mimic legitimate ones and deceive unsuspecting users into divulging sensitive information. Phishing attacks are the easiest way to acquire sensitive data from unsuspecting individuals. Phishers aim to obtain critical information like login credentials, passwords, and bank account details. By exploiting the user's vulnerability, hackers can gain access to a plethora of information, including social media passwords, company financial reports, and online transaction details. The objective of this study is to detect phishing URLs and determine the most effective machine learning method based on precision, false-positive, and false-negative rates.

I. INTRODUCTION

Phishing is a technique employed to deceive an individual via an electronic medium in order to acquire sensitive data, such as usernames, passwords, and credit card numbers. The approach often involves persuading customers to enter their personal information on a fake website that mimics the appearance and feel of the authentic one, typically achieved through email spoofing or instant messaging. This illicit practice has become one of the most perilous and damaging criminal activities proliferating online. In recent years, internet users who rely on online services have become increasingly susceptible to phishing attacks.

To gather sensitive information, scammers create unauthorized replicas of genuine websites and emails, often from financial institutions or other companies that handle financial data. They use the language and logos of the real organization to fabricate the email. The rapid growth of the internet as a communication platform is partially due to the way websites are designed, which also enables the misuse of trademarks, trade names, and other corporate identifiers that consumers rely on for identification purposes.

The forged emails are then disseminated to as many individuals as possible with the intention of deceiving them. When recipients receive these emails or click on a link within them, they are directed to a counterfeit website that appears to be from the genuine company.

Due to various cyber threats such as identity theft, personal information theft, and financial losses, internet users are vulnerable. Therefore, using the internet at home or at work can be precarious. Users should be equipped with efficient analytical tools to identify and safeguard against privacy breaches, thereby reducing security risks. To enhance self-intervention at the moment of an attack, an information security management system founded on artificial intelligence should be employed to create effective systems.

II. Literature Survey

Phishing is a fraudulent technique that uses a counterfeit website to obtain personal information, money, or data. The best way to avoid exposure to such websites is to detect dangerous URLs in real-time. The identification of phishing websites is reliant on their domains, which typically involve low-level and upper-level domains, paths, and queries that must be registered. By utilizing distinctive characteristics obtained from the words contained within a URL, as well as query data from search engines such as Google and Yahoo, it is feasible to evaluate the current status of intra-URL relationships. These attributes additionally influence machine-learning-based classification models for the detection of phishing URLs using real datasets.

The phish-STORM approach utilized in this study focuses on real-time URL phishing as opposed to phishing material. To differentiate between phishing and non-phishing URLs, several relationships between the registration domain and the rest of the URL are taken into account. However, the use of certain commonly blacklisted URLs to identify phishing websites is ineffective since these websites only exist for a short period of time due to the practice of phishing. Phishing is the act of deceiving a company's clients into divulging their sensitive information through unethical means. It can also be defined as the deliberate use of aggressive tools such as spam to automatically target victims and collect their private information.

With numerous SMTP failures serving as exploitation channels for phishing websites, there is an increased availability of communication channels for the delivery of malicious messages. To address this, a novel feature extraction method for classification that utilizes heuristics has been proposed. The extracted characteristics have been categorized into various categories, including features that obfuscate URLs and features that rely on hyperlinks. The

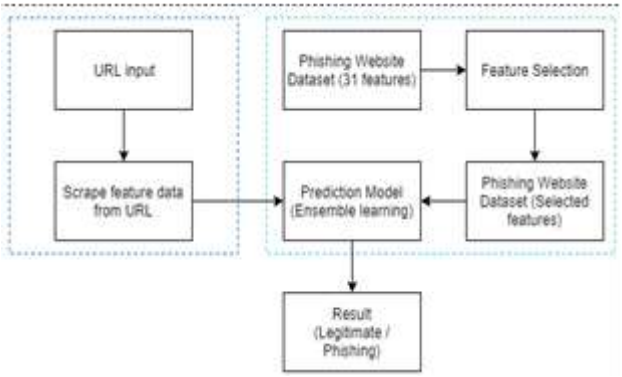
proposed approach achieves an accuracy of 92.5%. The only factors that affect the accuracy of this model are the number, quality, and feature extraction of the training data.

III. Methodology

Our project is based on a user-friendly website that provides software services to all users. This website is designed to help users determine whether a website is authentic or a phishing attempt. It features an engaging and responsive interface, making it easy for users to use and navigate.

To create this website, we utilized React, which is a popular front-end JavaScript toolkit that is free and open-source. React is specifically designed for building user interfaces using UI components. We made sure that the website is accessible and easy to use for all users, without any complications. As such, simplicity and ease-of-use were critical considerations during the website's development process.

Our website provides comprehensive information about the services we offer, as well as resources that educate users about unethical practices in the tech industry. The primary aim of the website is to help users identify and distinguish between authentic and fake websites. This will enable them to protect their personal information, including their email address, password, debit card number, credit card number, CVV number, bank account number, and more, from being misused by malicious actors. We believe that empowering users with this knowledge will help them to be more secure and confident in their online activities.



1. *Dataset Import*

Import a dataset from Kaggle.com that contains both genuine and phishing URLs, designated as "0" for trustworthy websites and "1" for malicious websites.

2. *Data preprocessing*

data preparation involves various steps such as data cleaning, data integration, data transformation, and data reduction. Data cleaning involves removing missing values, outliers, and noise from the data. Data integration involves merging data from different sources to create a consistent dataset. Data transformation involves transforming data to make it suitable for the analysis process, such as normalisation and scaling. Data reduction involves reducing the size of the dataset by selecting a subset of relevant features or samples to improve the efficiency of the analysis process. All of these steps are essential to ensure that the data used in the analysis is accurate, reliable, and suitable for the intended purpose.

3. *Trained ML Model*

Used Google Colab to train the model with features such as:

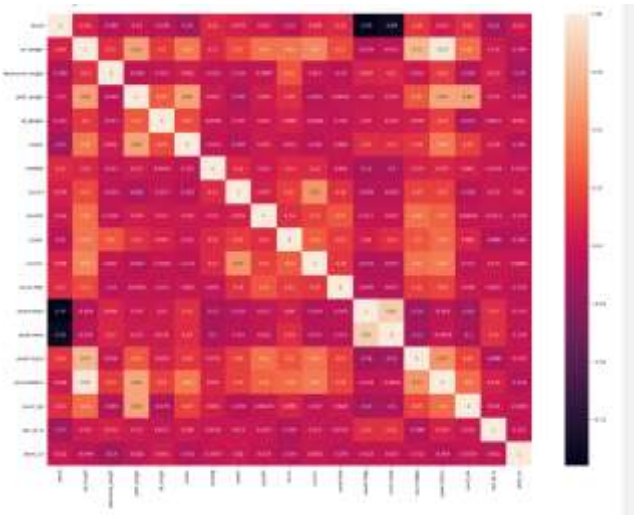
- URL redirection: The presence of "//" in the URL path indicates that the URL is a redirection to another website. If it is present, the feature is set to 1 to indicate that the URL is a potential phishing website. If it is not present, the feature is set to 0 to indicate that the URL is not a redirection.
- Length of Host name: The length of the host name is calculated and if it is greater than 25, the feature is set to 1, otherwise it is set to 0.
- Services "TinyURL": The @ sign is used by phishers to add a deceptive element to a URL. When the @ symbol is included in a URL, the browser ignores the content preceding it and jumps to the address following it. Therefore, if a URL contains an @ sign, the feature is set to 1 to indicate a potential risk of phishing. Otherwise, the feature is set to 0, indicating a lower likelihood of phishing activity.
- presence of @ symbol in URL: The presence of an IP address in the URL determines whether the feature is set to 1 or 0. Legitimate websites typically do not use IP addresses as part of their URL to load webpages, so if an IP address is detected, it may indicate an attempt by an attacker to obtain sensitive information.
- Presence of IP address in URL: If the IP address is included in the URL, the feature is set to 1, otherwise it is set to 0. A URL to download a webpage that contains an IP address is almost never used by reliable websites. If an IP address appears in a URL, the attacker may be attempting to steal sensitive information.

- Information submission to Email: The "mail ()" or "mailto:" methods can be used by phishers to collect user data and send it to their own email. If these functions are present in the URL, the feature is set to 1; otherwise, it is set to 0.
- Number of slash in URL: The number of slashes in a URL is another feature that can be used to determine whether a website is trustworthy or not. Typically, benign URLs have an average of five slashes. If the number of slashes in a URL is greater than five, the feature is set to 1, indicating a potentially malicious website. If the number of slashes is less than or equal to five, the feature is set to 0, indicating a potentially trustworthy website.
- Anchor URL: This feature was retrieved via crawling the anchor URL and its source code. The URL of the anchor is specified by the a> element. If the a> element has a maximum number of hyperlinks leading to other websites, the feature is set to 1; otherwise, it is set to 0.

MLP, or a multilayer perceptron, was employed. Another name for multi-layer perception is MLP. It is made up of thick, fully linked layers that may change any input dimension into the required dimension. A neural network with numerous layers is referred to as a multi-layer perception. In order to build a neural network, we join neurons so that some of their outputs are also their inputs.

4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a data analysis method that involves various techniques, many of which are visual and graphical. EDA helps to improve the understanding of a dataset by revealing its underlying structure, identifying important features, detecting outliers and anomalies, and testing assumptions. One common technique used in EDA is the creation of a heatmap, which can reveal patterns and relationships between variables in the dataset.



5. Develop API and host it on render.com

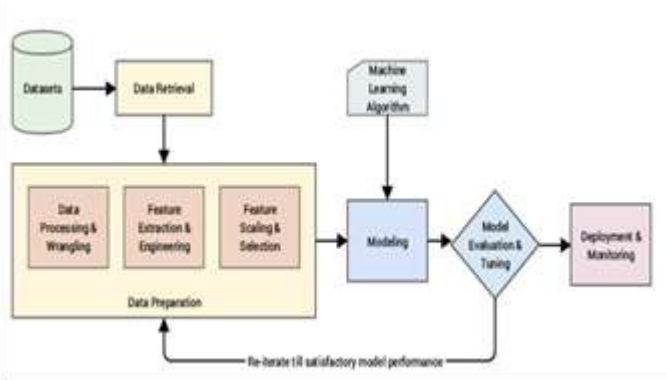
API or Application Programming Interface is a set of rules, protocols, and definitions that allow software components to interact with each other. In simpler terms, an API acts as an intermediary between two software applications, allowing them to communicate and share data seamlessly. The API that you are referring to has its source code available at https://github.com/prof-moriarty/fishyapi and it is hosted on Render.com. Render is a cloud-based platform that allows users to build and run their applications and websites. Render provides free TLS certificates, a global CDN, DDoS protection, private networks, and automatic deploys from Git. When you deploy your API on Render, it pulls the code from your Github repository and hosts it on their servers.

6. Develop a frontend with React and host it on Github Pages

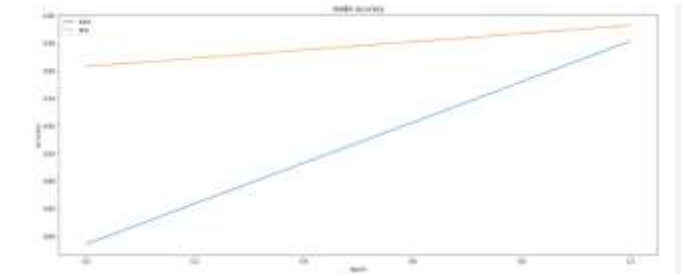
The user interface of the project is built using React, which is a popular and open-source JavaScript library for building interactive and reusable UI components. The frontend of the project can be accessed through Github Pages using the following link: https://prof-moriarty.github.io/fishy0/.

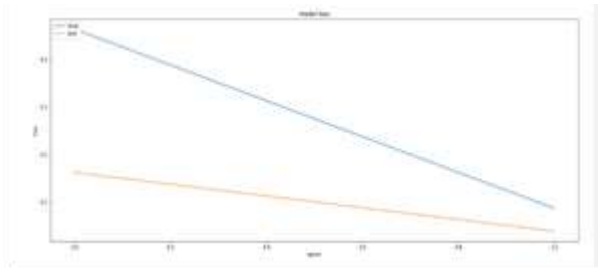
7. *Working*

The scanning process begins by entering a URL into the search field and clicking the Scan button. The entered URL is then sent to the API hosted on render.com, which applies a Multilayer Perceptron model to classify the URL as genuine or fraudulent. The model is trained to distinguish between real and phishing URLs. Once the URL is classified, a likelihood score is assigned to it, expressed as a percentage. If the score is above the 70% threshold, the URL is deemed highly likely to be a phishing attempt. On the other hand, if the score is below 70%, the URL is considered relatively safe.



8. *Comparison between different algorithm for accuracy*





IV. Conclusion

Phishing is a type of cybercrime that utilizes social engineering tactics in the computing industry to deceive users and steal their personal and sensitive information, such as login credentials and credit card details. Despite efforts to combat this issue, significant security challenges still exist, including the proliferation of botnets, spam email, and organized crime. To address these challenges, future work in this area may involve the development of unsupervised deep learning techniques that can extract knowledge from URLs. This research can be expanded to cover a larger network while respecting individual privacy rights.

The research suggests that using a classifier and a multilayer perceptron machine learning system can effectively identify web phishing. The study highlights that classifiers perform better when more features are included as training data, especially the most important characteristics. Currently, there are classifiers available that can detect phishing websites with high accuracy. However, scammers are continuously improving the URLs and designs of phishing websites to make them resemble legitimate websites. Therefore, it is crucial to continuously enhance the current features and add new ones to improve the accuracy of phishing detection.

V. References

[1] Joby James, Sandhya L., Ciza Thomas; "Detection of phishing URLS using machine learning techniques"; International Conference on Control Communication and Computing (ICCC); 2013; DOI:10.1109/ICCC.2013.6731669

[2] M Selvakumari, M Sowjanya, Sneha Das, S Padmavathi; "Phishing website detection using machine learning and deep learning techniques"; Journal of Physics Conference Series: 2021; DOI:10.1088/1742-6596/1916/1/012169

[3] Rishikesh Mahajan, Irfan Siddavatam,"Phishing Website Detection using Machine Learning Algorithms; International Journal of Computer Applications"; 2018; DOI:10.5120/ijca2018918026

[4] Arun Kulkarni, Leonard L. Brown; "Phishing Websites Detection using Machine Learning", International Journal of Advanced Computer Science and Applications (IJACSA), Volume 10, 2019; (DOI) 10.14569/IJACSA 2019.0100702

[5] Arathi Krishna V, Anusree A, Blessy Jose, Karthika Anilkumar, Ojus Thomas Lee, "Phishing Detection using Machine Leaming based URL Analysis: A Survey, National Conference on Novel & Challenging Issues and Recent Innovations in Engineering and Information Sciences (NCREIS); 2021, DOI: 10.17577/1JERTCONV9IS13033

[6] J. Kumar, A. Santhanavijayan, B.Janet, B. Rajendran and B. S. Bindhumadhava, "Phishing Website Classification and Detection Using Machine Learning," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1–6, 10.1109/ICCCI48352.2020.9104161.

[7] Hassan Y.A. and Abdelfettah B, "Using case- based reasoning for phishing detection", Procedia Computer Science, vol. 109, 2017, pp. 281–288.

[8] Rao RS, Pais AR. Jail-Phish: An improved search engine based phishing detection system. Computers & Security. 2019 Jun 1;83:246–67.