SPOOFGUARD-UNVEILING DECEPTIVE ONLINE PLATFORMS

Sushma V[1], Vignesh Srinivasan S[2], Chandraprabha K[3]

[1] Student, Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India
[2] Student, Computer Science and Business System, Bannari Amman Institute of Technology, Tamil Nadu, India
[3] Associate Professor, Information Technology, Bannari Amman Institute of Technology, Tamil Nadu, India

ABSTRACT

Phishing is a cyber-attack used by hackers and malicious organizations to steal the credentials like username, passwords, financial data, and bank details of individuals. They achieve this by pretending to be a legitimate service provider of a popular brand. Phishers create a fake website or clone of a trending website. They try to make the URL of the website like the legitimate one. The user might not notice the minute differences in the website and become the victim of this attack. Though phishing attacks are happening over a long period of time, it is still active and successful. The reason is lack of awareness among the people about the phishing attack. We propose a method that demonstrates the effectiveness of using PSO-based feature weighting to improve the detection of phishing websites and develop that into a Chrome extension which warns the user when the user enters a malicious website. This Chrome extension is designed to enhance users' online safety by identifying and flagging potential phishing websites in real-time. Leveraging advanced machine learning algorithms and PSO, it analyzes various features of web pages, including URL structure, content, and behaviour, to assess the likelihood of a webpage being a phishing site. PSO optimizes the machine learning model's parameters, enhancing its accuracy and adaptability. By incorporating PSO, this extension provides users with instant warnings and prompts them to exercise caution when encountering suspicious websites. This optimization technique allows the machine learning model to adapt and learn from evolving phishing tactics, ensuring its effectiveness against zero- hour attacks. Moreover, its lightweight and user-friendly design seamlessly integrates into users' browsing experiences without causing disruptions. This research contributes to the ongoing efforts to enhance online security and protect users from falling victim to phishing scams.

Keyword: - *Phishing, PSO, Swarm Intelligence, Chrome Extension, URL features, Machine Learning, Random Forest Classifier*

---

1. INTRODUCTION

With the expanding utilize of the web and advanced stages, the hazard of online assaults such as phishing has too increased significantly. Phishing could be a shape of online assault where cybercriminals make fake websites or emails that imitate true blue ones to deceive clients into giving touchy data as appeared in Fig. 1. This secret information has the potential to be abused for illegal activities such as financial extortion or information robbery. This approach to defending against such attacks has never been more crucial, and this is where our model for locating and anticipating phishing websites with machine learning comes in Machine learning is a collection of artificial intelligence that allows systems to memorize and improve their execution without human intervention. This could be referred to as machine learning. It includes making calculations competent of identifying information patterns and utilizing them to form expectations. Within the context of our demonstrate for discovery and anticipation of phishing websites, machine learning calculations are utilized to recognize designs within the data associated with websites, such as their URLs, HTML labels, and content substance. By utilizing these designs, the show can observe whether a site is veritable or misleading. Creating a demonstrate for identifying and avoiding

phishing websites utilizing machine learning is crucial in today's computerized scene. Phishing assaults are getting to be progressively modern, and conventional strategies of detecting and avoiding such assaults are now not sufficient.

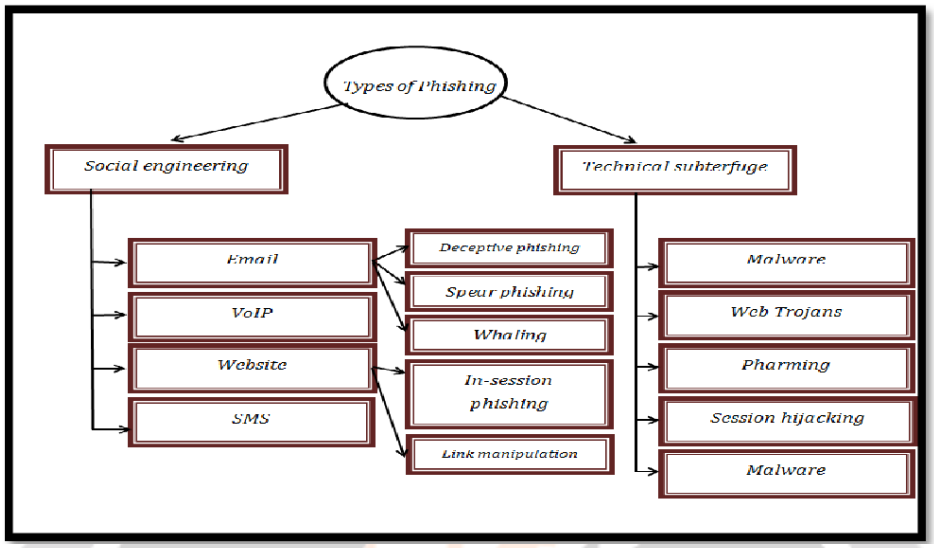

Fig -1: Types of Phishing

2. LITERATURE SURVEY

We have studied many papers so far and we have come across the different techniques and technologies which are used to develop the system and software to avoid or detect the phishing. The APWG recorded a staggering 1,025,968 phishing attacks during Q1 2022, marking a significant milestone as it was the highest quarterly total ever recorded, surpassing one million attacks for the first time in history. The study also revealed a notable 7% increase in credential theft phishing attacks targeting enterprise users, highlighting the growing risk to organizations. While most industries experienced a decrease in the number of ransomware attacks, the Financial Services sector bucked this trend, witnessing a concerning 35% increase in attacks during the first quarter of 2022.

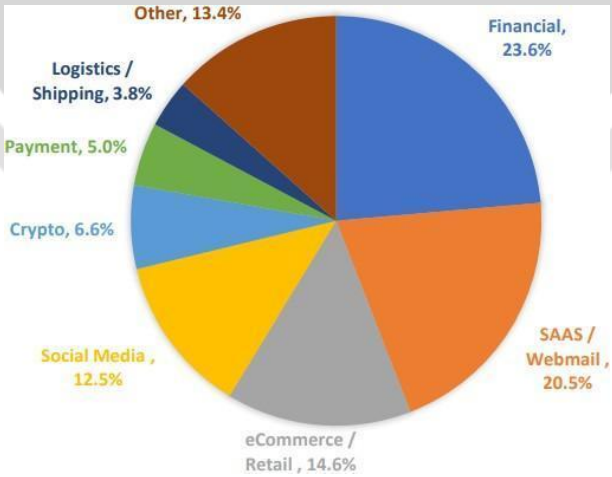

Fig -2: Most targeted industries

In Zou Futai, Gang Yuxiang, Pei Bei, Pan Li, Li Linsen "Web Phishing Detection Based on Graph Mining," 2016 in IEEE; it uses network Phishing Detection database mining methodology. This will detect any potential phishing attempts that the URL review is unable to detect. makes use of the user-website browsing configuration. to obtain the information gathered from actual traffic on a large ISP. A generic address is selected from the ISP's address database, but a unique AD is assigned to the client customer. As a result, we establish a visit relationship graph with

the AD and URL, which we refer to as the AD-URL graph. The reciprocal actions of the graph identify the phishing website.

In Nick Williams, Shujun Li "Simulating Human detection of phishing websites: An investigation into the applicability of ACT-R cognitive behavior architecture model", 2017 in IEEE; They proposed a framework for dissecting the idea of cognitive-behavioural architecture known as ACT-R. Analyze the neurological processes that go into determining a website's legitimacy by simulating its HTTPS padlock secure predictor characteristics, in particular. Further research to more accurately reflect the spectrum of human security awareness and activities in an ACT-R system may yield deeper insights into how to best integrate technology and human protection to lower the likelihood of phishing attacks for users. ACT-R has good modeling capabilities for the phishing use case.

R. Zieni at. el., 2023 intentionally compiled an extensive collection of pertinent papers, encompassing three primary categories of detection methods: list-based, similarity-based, and machine learning-based approaches. They thoroughly examined the strengths and weaknesses of these methods in their discussion.They have concluded that machine- learning based method is efficient due to their proficiency in identifying zero-hour attacks and effectively handling newly emerged phishing web pages. They also addressed the growing utilization of URL shortening services by malicious attackers affecting machine learning-based methods, as many conventional URL features recommended in the literature lose their significance in this context, leading to a potential failure in detection mechanisms.

3. METHODOLOGY

The development of a model for discovery and prevention of phishing websites using machine literacy requires a well-defined methodology. The methodology outlines the way involved in creating the model, from data collection to model deployment. This section discusses the methodology used in the development of our model

3.1 Dataset

1. Abnormal URL: Checks whether URL is without hostname.
2. '@' Symbol: Checks whether URL contains '@' symbol.
3. Sub domain: Calculates the number of sub domain on the basis of dots present in URL
4. URL requests: Calculates percentage of number of times URL is requested
5. Shortening services: Checks whether the URL is too short.
6. 'HTTPS' token: Checks whether the domain has HTTPS token or not.
7. Server from handler (SFH): Checks whether SFH contain "about: blank" or "Is empty"
8. URL with anchor: Calculates the percentage of anchor URL
9. Tag containing links: Calculates the percentage of links in 'Script', 'Meta' and 'link'.
10. Iframe: Checks whether webpage makes use of iframe.
11. IP Address: Checks whether URL contains IP Address.
12. Length of URL: Calculates and checks whether the length of URL is too big.
13. Double slash forwarding: Checks whether the '//' in URL is within 7th character.
14. Non-standard ports: Checks whether the port number has preferred status
15. Prefix and suffixes: Checks whether domain contains '_'.
16. Favicon: Checks whether favicon is retrieved by external or internal source
17. SSL final certificate: Checks if URL is using https by trusted providers and certificate age.
18. Age of domain: Calculates the age of Domain.
19. DNS record: Checks whether the domain is with or without DNS record.

20. Links pointing to page: Calculates how many links are pointing to the Webpage.

21. Domain registration length: Checks the expiry date of Domain

3.2 Data Collection

The first thing to do is collect data. Gathering data entails compiling a sizable inventory of trustworthy and phony websites. The dataset must to be substantial enough to encompass the diversity among the various phishing assault kinds. Similar to web scraping or using datasets, there are several methods for gathering data. For this concept, a dataset from Phishtank was utilized.

3.3 Data Preprocessing

The next stage is to preprocess the data after it has been collected. Data pre-processing includes preparing the data for analysis by cleaning and converting it. Because it guarantees that the data is of the highest quality and appropriate for machine literacy algorithms, this stage is crucial. Preparing the data for machine learning algorithms, eliminating duplicates, and handling missing values are all part of the pre-processing stage. Using a box plot and other techniques, the aforementioned dataset was examined for duplicates, missing values, null values, and other noise. The features in the aforementioned dataset that weren't very helpful in phishing detection were removed.

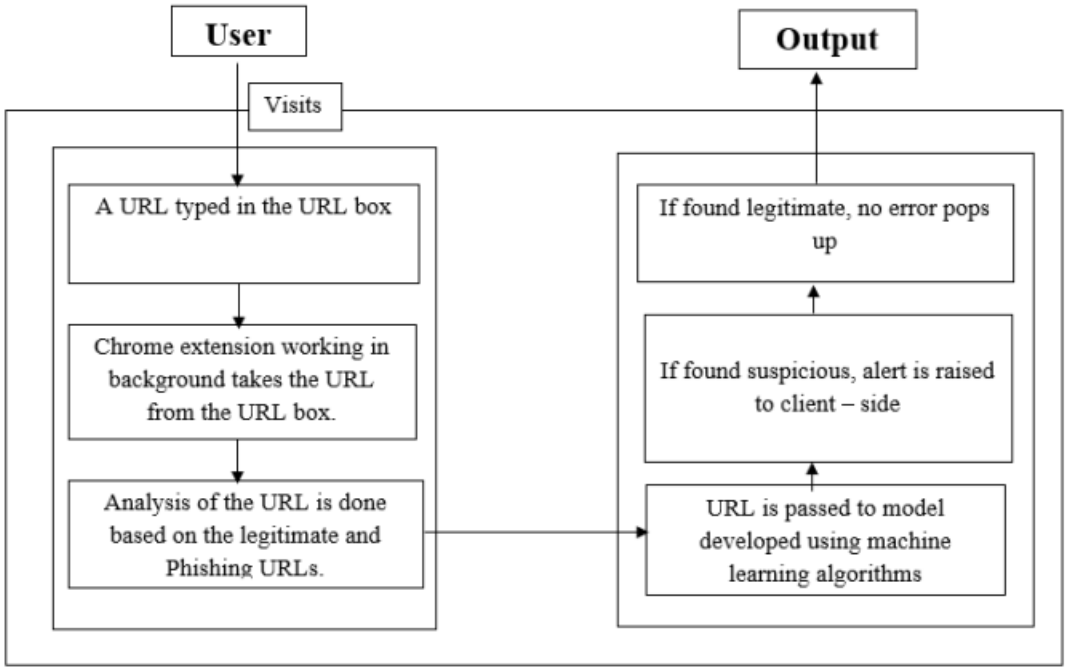

Fig -3: Design Methodology

3.4 Model Selection

The next stage is to select a machine literacy model by hand. Selecting a model entails deciding on a swish algorithm that can distinguish between trustworthy and phony websites. Numerous machine learning methods are available, including Random Forest, Decision Tree, Support Vector Machines, Gradient Boosting Classifier, K-Nearest Neighbors, AdaBoost Classifier, Logistic Regression and others. In this design, we used a random forest algorithm

For this project, we trained the model using various algorithms. For each the accuracy, F1-score, precision, recall was calculated. The algorithm which provided the best values was random forest. Hence, random forest was used for training and testing the model. We have trained and evaluated our model using 22 parameters dataset and various algorithms. From all the algorithms used Random Forest provided the best accuracy. For the implementation of extension, 16 parameters were used and single layer perceptron was used to detect phishing websites

3.5 Random Forest Algorithm

Based on the values of the randomly selected variable, several tree predictors are incorporated into random forests, In actuality, every forest tree has the same distribution. For the majority of data, the Random Forest algorithm can provide an accurate class prediction. However, there are a few errors that trees occasionally make as well. We therefore choose to observe the class on the poll results by conducting the vote for each observation in order to help the. determine the outcome with greater accuracy.
In our proposed system we use this algorithm to predict if the web URLs are phishing or not. We use this algorithm for the classification of features on the data to predict the class from it. It consists of multiple instances with multiple outcomes. On that basis, the different decision trees are formed according to different features and outcomes and with the same features with multiple outcomes. The final result is based on the majority pole of the multiple decision trees altogether and the class prediction is based on it. In addition, an internal unbiased calculation of the generalization error is produced during the forest construction process. In fact, incomplete data can be well calculated. The loss of reproducibility is a big downside to wild forests, as the method of creating the forest is arbitrary. Furthermore, it is difficult to understand the final model and the resulting effects, since it includes several different variables decisions trees

3.6 Model Training

After concluding the machine learning algorithm, the coming step is to train the model. Model training involves feeding the model with the pulled features and their corresponding labels. The model is trained by learning from the data and adjusting its parameters to minimize errors. According to this, data is divided into a training set and a validation set. The training set is utilized to train the model, whereas the validation set is used to assess the performance of the model. In this particular study, 80% of the data from the aforementioned dataset was employed to train the model.

4. CONCLUSIONS

4.1 Model Evaluation

After training the model, the coming step is to estimate its performance. To evaluate the effectiveness of the model, it is tested on a separate test set that has not been used in training. The performance of the model is then measured using various metrics such as sensitivity, accuracy, and recall. These metrics help in determining the model's capability to detect and prevent phishing attacks. 20% of data is used for testing the model. A clear technique is necessary for the creation of a model that uses machine literacy to identify and block phishing websites. Data collection, preprocessing, feature lodging, machine knowledge model conclusion, model training, model performance evaluation, and model planting are all steps in the methodology. We can create a model that offers real-time defense against phishing attempts by using this process.

4.2 Results

The extension shows an alert to the user when a phishing website is visited to save the user from any harm. For this to work the extension needs to be added in the browser by the user. Extension is implemented using single layer perceptron in deep learning after evaluating the accuracy of deep learning and various machine learning models. The websites are detected as phishing based on various URL features. When any website is visited the URL of the website is extracted by extension, the extension then processes the URL to detect it as safe or phishing, this processing is done by JavaScript file(content.js).

5. FUTURE SCOPE

Deep learning methods such as Generative Adversarial Network (GAN) and Recurrent Neural Network (RNN) can be employed in the future. In addition, we want to inform the original users about their phishing websites. For instance, if a phishing website similar to www.microsoft.com is made, say www.microsooft.com, we want to let the original website creator know about the phishing website that has been made under their name. Additionally, emotional analysis and web scraping can be used to identify the kind of websites that display phony things, take money, then disappear without providing any goods in return.

6. REFERENCES

[1]. R. Zieni., L. Massari., & M. C. Calzarossa. (2020). Phishing or Not Phishing, A Survey on the Detection of Phishing Websites.
[2]. APWG (2022). Phishing Activity Trends Report 1st Quarter.
[3]. S. Haruta., H. Asahina., & I. Sasase. (2017). Visual Similarity-Based Phishing Detection Scheme Using Image and CSS with Target Website Finder. IEEE Global Communications Conference. 1-6.
[4]. M. D. Bhagwat., P. H. Patil., & T. S. Vishawanath. (2021). A Methodical Overview on Detection, Identification and Proactive Prevention of Phishing Websites. Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV). 1505-1508.