



Research**PREVENTION OF PHISHING**¹MR.K.RAJA,² B. JEEVITHA, D. ESHWAR,⁴ J. MADHUNESHWARI,⁵ K. SAI CHARAN¹Assistant Professor, Department of DS, Sri Indu College Of Engineering & Technology.^{2,3,4,5}U.G.Scholar, Department of DS, Sri Indu College Of Engineering & Technology, Hyderabad**Abstract:**

Phishing is a common cyberattack in which fake websites are used to deceive individuals into revealing their personal information. Phishing URLs and fraudulent webpages are often designed to steal sensitive data such as usernames, passwords, and online banking details. Attackers create websites that visually and structurally resemble legitimate websites in order to trick users into trusting them. As technology has advanced, phishing techniques have also become more sophisticated, making it increasingly difficult for users to identify such threats. To address this issue, anti-phishing systems can be developed to detect and prevent phishing attacks. Machine learning has emerged as a powerful approach for identifying phishing activities by analyzing patterns and characteristics of suspicious URLs and websites. The proposed system explores the use of machine learning techniques to improve the detection of phishing attacks. By studying different features of web links and webpage structures, the system aims to accurately identify phishing websites and help protect users from potential cyber threats.

Keywords: Phishing, Phishing Websites, Detection, Machine Learning**I. INTRODUCTION**

Phishing imitates the features and options of emails and makes them seem identical to the real thing. It resembles the real supply quite closely. The customer believes that this email is from a legitimate job or business. This forces the user to click on the links provided in the phishing email and visit the phishing website. These phishing websites were developed to imitate the design of a clever website. The phishers coerce people into listing up their private information by sending false messages, asking them to confirm their account, and other means so that they may list up the information they want to use against them. They come up with strategies that prevent users from constantly having an option but to visit their fake website. The most dangerous criminal activity in the online world is phishing. Since the majority of users log on to access the services offered by governmental and financial institutions, phishing attempts have significantly increased over the past several years. Phishers started using this as a lucrative business to make money. Phishing may be illegal, but those that engage in it do so because they find it to be incredibly trustworthy, efficient, and doesn't cost anything. The email identity of someone may actually be obtained by phishing, therefore it's extremely sincere to check for it out every day and send an email to everyone since it's publicly available everywhere. These attackers have a terrible lack of resources, making it difficult for them to quickly and effectively advance vital knowledge. The phishing scams have an impact on fraud, statistics loss, malware infections, etc. The vital information of a user, such as their password, one-time password, credit/debit card numbers, CVV, sensitive company information, medical knowledge, secret information, etc., is information in which those cybercriminals are interested. These thieves frequently also gather information that will give them direct access to social media. There are several programs, methods, and algorithms that are used to identify phishing. These comprise utilized at the academic and corporate levels. Take for instance that to cover the original domain selection the phishing assaulter would feel horribly long and complicated name of the domain. A phishing address and also the parallel internet page have numerous characteristics that may be unique from the address. This is frequently horribly obvious

II. LITERATURE SURVEY

In [1] Various machine learning algorithms logistic regression, decision tree classifier, random forest classifier, AdaBoost, gradient boosting classifier for the phishing detection.

In this paper logistic regression [2], Gaussian Naïve Bayes and Random Forest were been Proposed Proposed[3], a

This article can be downloaded from <https://ijerst.org/index.php/ijerst>

machine learning-based phishing detection system by using eight different algorithms to analyze the URLs, and three different datasets to compare the results with other works.

In [4] Review was made With the huge number of phishing emails or messages received every day, companies or individuals are not able to detect all of them, where different reviews were given for detection of phishing attack, by using machine learning.

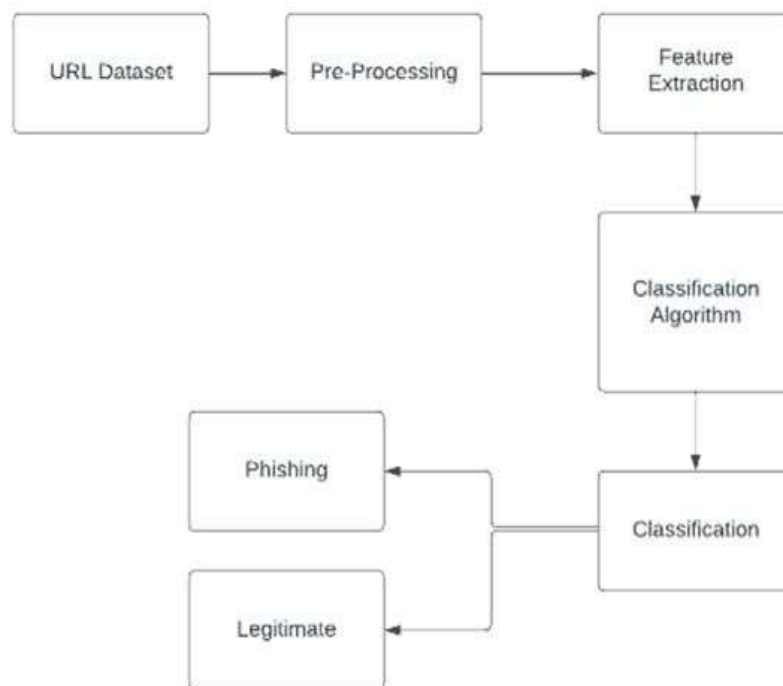
Proposed [5] three approaches for detecting phishing websites. First is by analyzing various features of URL , second is by checking legitimacy of website by knowing where the website is being hosted and who are managing it, the third approach uses visual appearance based analysis for checking genuineness of website.

In [6] Proposed System provides an approach to detecting phishing email attacks using an analysis of linguistic communication and machine learning. It is accustomed search the text's syntax to detect malicious intent. A natural language processing (NLP) technique is employed in conjunction with a predicate to decode each sentence and identifies the semantic jobs of words within the sentence. Computer supervised learning is employed to come up with the blacklist of malicious pairs.

III. OBJECTIVE OF SYSTEM

The main purpose of the proposed system is to detect the phishing websites who are trying to get access to sensitive data or by creating the fake website and trying to get access of user's personal credentials. We are using machine learning algorithm to safe guard the sensitive data and detect the phishing website who are trying to gain access on sensitive data.

IV. IMPELNTATION OF MODULE



The Figure a: Block Diagram

The proposed system provides a dataset of legal and phishing URLs, which is pre-processed to put the data in a functional format for analysis. The features revolve on several characteristics of phishing websites that have been utilized to separate them from genuine ones. The characteristics and values of each type of phishing are clearly stated. For each URL, the desired features are retrieved, and valid stages of inputs are located. Each risk associated with a phishing website is then given one of these values. **Methodology**

The proposed phishing detection system is designed to identify whether a given URL or website is legitimate or malicious by using machine learning techniques. The methodology consists of several stages, starting from data collection and ending with prediction of phishing websites.

1. Data Collection

A dataset containing both legitimate and phishing URLs is collected from reliable sources. The dataset includes different website and URL-based features that help distinguish phishing websites from genuine ones.

Examples of features:

- URL length
- Presence of @ symbol
- Use of - in domain
- Number of subdomains
- Presence of HTTPS
- Age of domain
- Use of IP address instead of domain name
- Redirection using //
- Suspicious prefix or suffix in URL

2. Data Preprocessing

The collected dataset is cleaned and prepared before training the model.

This step includes:

- Removing duplicate entries
- Handling missing values
- Converting categorical values into numerical form
- Normalizing or scaling data if needed
- Labeling data as:
 - 0 for legitimate website
 - 1 for phishing website

3. Feature Extraction

Important characteristics are extracted from the URL and webpage content. These features are used as input for the machine learning model.

Feature categories include:

- **URL-based features**
- **Domain-based features**
- **HTML and JavaScript-based features**
- **Security-related features**

These extracted features form the input vector for classification.

4. Feature Selection

Not all extracted features contribute equally to the prediction. Feature selection is performed to identify the most relevant features and improve the efficiency and accuracy of the model.

Common techniques:

- Correlation analysis
- Information gain
- Chi-square test
- Recursive feature elimination

5. Model Training

The processed dataset is divided into:

- **Training set** for learning
- **Testing set** for evaluation

A machine learning classification algorithm is then trained on the training data. Common algorithms used for phishing detection are:

- Naïve Bayes
- Decision Tree
- Random Forest
- Support Vector Machine
- K-Nearest Neighbors

- Logistic Regression

Among these, the best-performing algorithm can be selected based on accuracy and other evaluation measures.

6. Model Testing and Evaluation

After training, the model is tested using unseen data. Its performance is evaluated using metrics such as:

- Accuracy
- Precision
- Recall
- F1-Score
- Confusion Matrix

These metrics help determine how effectively the model can detect phishing websites.

7. Prediction Phase

When a user enters a URL, the system extracts the same set of features from that URL and passes them to the trained machine learning model. The model then predicts whether the URL is:

- Legitimate
- Phishing

8. Result Display

The final result is displayed to the user, warning them if the URL is suspicious or confirming if it is safe.

The double no 0 and 1 that appears the characteristic is present or not is used to address the phishing features

V. CONCLUSION

The education and knowledge of phishing attacks is the most crucial step in protecting the user. Internet users must be familiar with any security recommendations provided by experts. Additionally, users should be taught not to heedlessly click on links that direct them to websites where they must submit sensitive data. It is crucial to verify the URL before to visiting the website. It has developed into a significant network security issue, resulting in billions of dollars in losses for both customers and e-commerce businesses. Phishing has also, and probably more importantly, contributed to the mistrust and allure of e-commerce among regular consumers. The education and knowledge of phishing attacks is the most crucial step in protecting the user. Internet users need to be mindful of any security advice provided by professionals. Additionally, users should be taught not to heedlessly click on links that direct them to websites where they must submit sensitive data. Before entering the website in the suggested system, it is important to check the URL

REFERENCES

- [1]. Lakshmanarao and P.SuryaPrabhakara Rao, "Phishing website detection using novel machine learning fusion approach", IEEE 2021
- [2]. Jitendra Kumar and A. Santhanavijayan, "Phishing Website Classification and Detection Using Machine Learning", International Conference on Computer Communication and Informatics, 2020
- [3]. Mehmet Korkmaz and Ozgur Koray Sahingoz, "Detection of Phishing Websites by Using Machine Learning-Based URL Analysis", IEEE 2020 Charu
- [4]. Singh, "Phishing Website Detection Based on Machine Learning: A Survey", IEEE 2020
- [5]. Vaibhav Patil and Pritesh Thakkar, "Detection and Prevention of Phishing Websites using Machine Learning Approach", IEEE 2018
- [6]. T. Peng, I. Harris, and Y. Sawa, "Detecting Phishing Attacks Using Natural Language Processing and Machine Learning," Proc. - 12th IEEE Int. Conf. Semant. Comput. ICSC 2018, vol. 2018-Janua, pp. 300–301, 2018