IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PHISHING WEBSITE DETECTION SYSTEM USING MACHINE LEARNING ALGORITHMS

[1]Guru Prasaath M, [2]Imran Khan I, [3]Muthu Lingam K, [4]Dr R Ramya

[1]Student at SRMIST, Ramapuram, [2]Student at SRMIST, Ramapuram,
[3]Student at SRMIST, Ramapuram, [4]Asst. Prof. CSE at SRMIST,Ramapuram.
[1]Department of Computer Science and Engineering,
[1]SRMIST, Ramapuram, Chennai, India.

Abstract: Phishing is an internet scam in which an attacker sends out fake websites or messages that look to come from a trusted source. Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password, and bank account details. Many researchers have spent decades creating unique approaches to automatically detect phishing websites. Machine learning based phishing detection systems can provide a more effective and efficient way to detect phishing attacks. It also eliminates the disadvantages of the previous method. The goal of this research is to use the dataset collected to train ML models. Traditional methods of detecting phishing attacks are not always effective because attackers can quickly create new URLs or use other tactics to evade detection. The ultimate goal is to use machine learning algorithms and OCR Technology to phishing detection system. Machine learning algorithms make the system more accurate and efficient. Using Optical Character Recognition (OCR) in a phishing detection system can enhance its capabilities by extracting text from images, which is often used in phishing attacks. These prevent users from falling victim to them.

Index Terms - Machine Learning, Classifiers, SVM, Decision Tree

I. INTRODUCTION

Phishing attack is a simplest way to obtain sensitive information and it has become the most serious problem, harming individuals, corporations, and even entire countries. The availability of multiple services such as online banking, entertainment, education, software downloading, and social networking has accelerated the Web's evolution in recent years. Now, Internet is the place where everyone shares a data and collects the data. These make the attackers to steel data by creating a phishing website.There are multiple websites who ask user to provide sensitive data such as username, password or credit card details etc. often for malicious reasons. This type of websites is known as phishing website. In order to detect and predict phishing website, we proposed an intelligent, flexible and effective system that is based on machine learning algorithms. Phishing detection system using machine learning are designed to identify and prevent phishing attacks, These systems use machine learning algorithms to analyze various features of emails, URLs or other digital content to determine whether they are legitimate or fraudulent. Phishing attacks are a significant threat to individuals and organizations, and they can result in financial losses, identity theft, and other serious consequences.

II. RELATED WORKS

Armando Maule and Marco Prandini [1] introduce a machine learning approach for phishing detection based on feature selection and classification techniques.

T.A. Murugesh, S. Vijayarani, and T. Santhosh Kumar [2] explores various machine learning algorithms for phishing website detection and evaluates their performance.

K. Saravanakumar and K. Arun [3] focuses on detecting phishing URLs shared on Twitter using machine learning techniques.

M. Firdhous and R. Perera [4] approaches that combines multiple machine learning algorithms for improved phishing detection.

S. A. Mahmood, B. Hu, and L. Hu [5] provides an overview of various machine learning techniques applied to phishing detection.

III. EXISTING SYSTEM

Email Filtering: Many email services and clients use filters to identify and block phishing emails based on known patterns, sender reputation, and content analysis.

URL Analysis: Some systems analyse URLs in emails or messages to check if they lead to suspicious or known phishing websites.

Machine Learning: Phishing detection systems often employ machine learning algorithms to detect unusual email patterns, content, or sender behaviour.

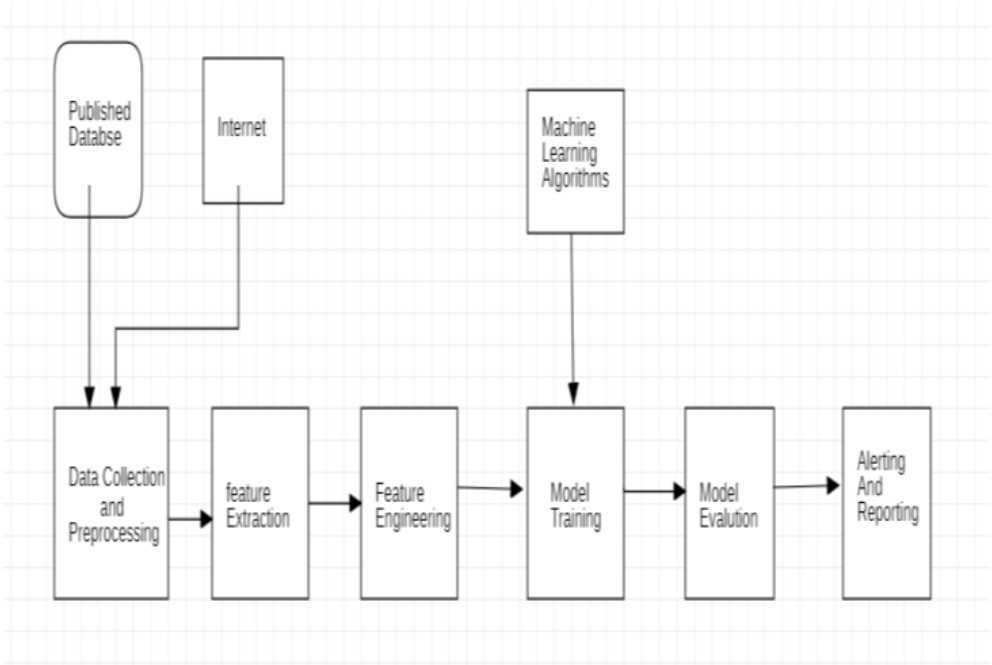
Blacklists: Maintaining lists of known phishing websites or malicious IP addresses helps identify and block phishing attempts.

Domain Verification: Some systems verify the legitimacy of a sender's domain by checking DNS records and digital signatures.

Multi-Factor Authentication (MFA): MFA adds an extra layer of security, making it harder for attackers to compromise accounts.

IV. PROPOSED WORK

The most frequent type of phishing assault, in which a cybercriminal impersonates a well-known institution, domain, or organization to acquire sensitive personal information from the victim, such as login credentials, passwords, bank account information, credit card information, and so on. Emails containing malicious URLs in this sort of phishing email contain a lot of personalization information about the potential victim. Phishers increasingly use images instead of text to evade detection. Traditional text analyze may miss these types of attacks. To overcome that we are proposing to integrate OCR (Optical Character Recognition) technology can enhance the system ability to analyze and detect phishing attempts that involve text with images



The modules in this project are Data Collection and Preprocessing, Feature Extraction, Feature Engineering, Machine Learning Model, Model Evaluation, Real-time Scoring and Classification, Alerting and Reporting, Scalability and Redundancy, Security and Access Control, Deployment and Integration, Documentation and Training.

V. DATASET

We collected the datasets from the open-source platform called Phishing tank. The dataset that was collected was in csv format. There are 18 columns in the dataset, and we transformed the dataset by applying data pre-processing technique. To see the features in the data we used few of the data frame methods for familiarizing. For visualization, and to see how the data is distributed and how features are related to one another, a few plots and graphs are given. The Domain column has no bearing on the training of a machine learning model. We now have 16 features and a target column. The recovered features of the legitimate and phishing URL datasets are simply concatenated in the feature extraction file, with no shuffling. We need to shuffle the data to balance out the distribution while breaking it into training and testing sets. This also eliminates the possibility of over fitting during model training.

VI. MACHINE LEARNING CLASSIFIERS

Decision Tree Classifier: For classification and regression applications, decision trees are commonly used models. They basically learn a hierarchy of if/else questions that leads to a choice. Learning a decision tree is memorizing the sequence of if/else questions that leads to the correct answer in the shortest amount of time. The method runs through all potential tests to discover the one that is most informative about the target variable to build a tree.

SVM Classifier: Support Vector Machine (SVM) is a machine learning algorithm used for both classification and regression tasks. It's particularly effective for classification problems and is known for its ability to handle high-dimensional data and complex decision boundaries.

Naive Bayes Classifier: a probabilistic machine learning algorithm used for classification tasks. It's based on Bayes' theorem and makes an assumption that features are independent, which is why it's called "naive." Despite this simplifying assumption, Naive Bayes can work surprisingly well for various types of data, especially in text classification and spam detection. Naive Bayes is known for its simplicity and speed, making it a popular choice for quick and efficient classification tasks, especially when dealing with large datasets and high dimensions.

Random Forest Classifier: Random forests are one of the most extensively used machine learning approaches for regression and classification. A random forest is just a collection of decision trees, each somewhat different from the others. The notion behind random forests is that while each tree may do a decent job of predicting, it will almost certainly over fit on some data. They are incredibly powerful, frequently operate effectively without a lot of parameters adjusting, and don't require data scalability.

VII. RESULT

When evaluating the classifier's behavior on the testing dataset, there were four statistical numbers There are Various performance measures namely False Positive Rate, False Negative Rate, Accuracy of the system is estimated. The basic count values such as True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) are used by these measures.

False Positive Rate (FPR): The percentage of cases where an image was classified to normal images, but in fact it did not.

$$FPR=FP/FN+TN$$

False Negative Rate (FNR): The percentage of cases where an image was classified to abnormal images, but in fact it did.

$$FNR=FN/FN+TN$$

Accuracy: We can compute the measure of accuracy from the measures of FPR and FNR as specified below.

$$Accuracy=TP+TN/TP+TN+FP+TN*100$$

Precision: The percentage of correctly identified positive data points among those predicted as positive by the model. The number of false-positive cases (FP) reflects the false warning rate. In real-time phishing detection systems, this directly affects the user experience and trustworthiness

$$Precision = TP/TP + FP$$

Recall: The recall is the portion of positive data points labeled as such by the model among all truly positive data points.

$$Recall = TP /TP + FN$$

Dateset Split Ratio (Training /Testing)	Classifiers	Accuracy Score	False Negative Rate	False Positive Rate
50:50	Decision Tree Classifier	96.71	3.69	2.93
	Random Forest Classifier	96.72	3.69	2.91
	SVM Classifier	96.40	5.26	2.08
	Decision Tree Classifier	96.80	3.43	2.99
70:30	Random Forest Classifier	96.84	3.35	2.98
	SVM Classifier	96.40	5.13	2.17
90:10	Decision Tree Classifier	97.11	3.18	2.66
	Random Forest Classifier	97.14	3.14	2.61
	SVM Classifier	96.51	4.73	2.34

Result shows that machine learning algorithms gives better detection accuracy which is 97.14 with lowest false negative rate than decision tree and support vector machine algorithms. Result also shows that detection accuracy of phishing websites increases as more dataset used as training dataset. All classifiers perform well when 90% of data used as training dataset.

VIII. REFERENCE

1. Kumar Dutta, A. (2019). Detection of Phishing Web as an Attack: A Comprehensive Analysis of Machine Learning Algorithms on Phishing Dataset. IOSR Journal of Engineering, 1-7. http://iosrjen.org/Papers/NCIRST-22/Volume%20-1/1,%2001-07.pdf
2. Singh, S., & Singh, S. (2017). Phishing Detection in E-mails using Machine Learning. ResearchGate. https://www.researchgate.net/publication/320257918_Phishing_Detection_in_E-mails_using_Machine_Learning
3. Kumar Dutta, A. (2021). Detecting phishing websites using machine learning technique. PMC - NCBI. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8504731/
4. Alzahrani, A. I., & Alzahrani, M. A. (2019). A predictive model for phishing detection. ScienceDirect.com. https://www.sciencedirect.com/science/article/pii/S1319157819304902
5. Singh, S., & Singh, S. (2022). Detection of Phishing Websites by Using Machine Learning-Based URL Analysis. International Journal of Research in Technology and Innovation, 1-8. https://www.ijrti.org/papers/IJRTI2207237.pdf
6. Salahdine, F., El Mrabet, Z., & Kaabouch, N. (2022). Phishing Attacks Detection: A Machine Learning-Based Approach. arXiv. https://arxiv.org/pdf/2201.10752.pdf