WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Phishing Website Detector Using Machine Learning

Prof. Abhishek P Nachankar[1], Swaraj Pawar[2], Tejaswini Potdukhe[3], Gourav Kayarkar[4], Aniket Kamble[5]*

[1]Assiatant Professor, Department of Computer Science & Engineering, KDK College of Engineering, Nagpur
[2,3,4,5]Students, Department Of Computer Science & Engineering, KDK College of Engineering, Nagpur
[1]Email: swarajgpawar@gmail.com [2]Email: tejaswinipotdukhe12@gmail.com

A B S T R A C T

Since mobile devices have become so common, there is a trend toward moving practically all offline activity online. Due to the anonymity of the Internet, this breaks several security laws even though it simplifies our daily lives. The simplest method for obtaining sensitive information from unwitting users is through phishing attacks. Phishers seek to get private data, including user-names, passwords, and bank account details. Cyber-security experts are searching for consistent and dependable methods of detecting phishing websites. In this research, numerous properties of both genuine and phishing URLs are extracted and analyzed in order to detect phishing URLs. Phishing websites can be recognized using decision trees, random forests, and support vector machine algorithms. This project ideally deals with machine learning technology to detect phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision trees, random forests, and support vector machines are algorithms used to identify phishing websites

Keywords: Phishing Attacks, Cybersecurity, Support vector, Algorithm, Extracting, Analyzing, Decision Tree, Random Forest.

1. Synopsis

With the proliferation of mobile devices in recent years, there is a tendency to move almost all real-world activities to the cyber world. Although this makes our daily life easier, it violates many security rules due to the anonymous nature of the Internet. Phishing attacks are the easiest way to get sensitive information from innocent users. Phishers aim to obtain sensitive information such as usernames, passwords, and bank account information. Cyber-security professionals are looking for reliable and consistent detection methods to detect phishing websites. This project deals with machine learning technology to detect phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision trees, random forests, and support vector machine algorithms are used to identify phishing websites in a two-step process of first visualizing and extracting features of URLs using python libraries and then training them into a model using Gradient Classifier Algorithm to predict real-time phishing websites.

2. Technical Keywords

2.1 Area of Project:

The purpose of this project is to provide an overview of the phishing website detection project, outlining the methodology, findings, and recommendations for improving the detection of malicious websites.

2.2 Technical Keyword :

The presented report includes the following technical keywords and abbreviations. Phishing, Website Analysis, Machine Learning, Data Mining, Feature Extraction, Classification Algorithms, URL Analysis, HTML Parsing, Blacklist Check.

Abbreviations- HTML, URL, SSL, TLS.

3. Introduction

3.1 Project Idea

Phishing attacks are the simplest way to obtain sensitive information from innocent users. The goal of the phishers is to acquire critical information like usernames, passwords, and bank account details. Everyone is now looking for trustworthy and steady detection techniques for phishing website detection. This project deals with machine learning technology for the detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Trees, random forests, and Support vector machines are the algorithms used to detect phishing websites. Support vector machines are the algorithms used to detect phishing websites.

3.2 Motivation of project

1.1 The main purpose of the project is to detect a website created by any phisher for hacking data from a user and making aware the user of such threats once detected. It proposes to prove much beneficial for users for safe browsing and keeping their data untouched by any phisher who is trying to use the user's credentials in illegal means.Working in ML Technology, the system is already proposed previously but this presented model can extract all the features and train the dataset for real time usage and also it is efficient then the previously proposed models.

3.3 Literature survey

Review of the papers, Description , Mathematical Terms,The Literature Survey of this project was available to us through various sources and from that we were able to pick a few so that we can understand what our project must include that will make it better then other previously proposed projects. Below mentioned are some of the major papers referred to by us in creating this project on "PHISHING WEBSITE DETECTOR USING ML". In order to get more information about the algorithms and technical concepts we browsed various internet sources, sat at libraries and read various documents to understand the current scenarios of phishing attacks and it's counter measures. After understanding the various terms, we were able to make this project a success.

4. Problem Statement

4.1 Problem statement

Nowadays Phishing has become a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. Phishing attacks are becoming successful because of lack of user awareness. Since phishing attacks exploit the weaknesses found in users, it is very difficult to mitigate them, but it is very important to enhance phishing detection techniques.

4.2 Goals and objectives

The main purpose of the project is to detect a website created by any phisher for hacking data from a user and making aware the user of such threats once detected. It proposes to prove much beneficial for users for safe browsing and keeping their data untouched by any phisher who is trying to use the user's credentials in illegal means.

Machine learning technology consists of many algorithms which requires past data to make a decision or prediction on future data. Using this technique, algorithms will analyze various blacklisted and legitimate URLs and their features to accurately detect the phishing websites including zero- hour phishing websites.

4.3 Statement of scope

The scope of this final year project is to develop a cybersecurity tool that specifically focuses on detecting and mitigating phishing websites. The objective is to create an efficient and reliable system that can identify and alert users about potential phishing attacks, helping them protect their sensitive information and reduce the risk of falling victim to online scams.

4.4 Hardware Resource Required

Sr. No.	Parameter	Minimum Requirement	Justification
1	CPU Speed	2 GHz	Considering Basic Devices
2	RAM	3 GB	RAM may vary on the system

5. Project Plan

It's important to note that project estimation can vary based on factors such as project complexity, team expertise, and available resources. It is advisable to involve relevant stakeholders, including team members and domain experts, to provide input and validate the estimates. Regularly review and update the project estimates as the project progresses to ensure accuracy and alignment with the project's evolving requirements..

Here are some key factors to consider for project estimation:

Data Collection and Pre-processing:

Estimate the time and effort needed to gather a diverse dataset of legitimate and phishing websites. Consider the complexity of data cleaning, normalization, and dataset splitting.

Feature Extraction and Selection:

Assess the complexity of identifying relevant features from website URLs, HTML content, SSL certificates, and email headers. Estimate the effort required to develop algorithms or techniques for feature extraction and selection.

Model Selection and Training:

Consider the time and effort needed for researching, evaluating, and selecting suitable machine learning algorithms. Estimate the training time required for the selected algorithms, considering the size and complexity of the dataset.

Model Evaluation and Performance Metrics:

Estimate the time required to evaluate the trained models using validation datasets and calculate performance metrics. Consider the complexity of comparing different models and selecting the best-performing one(s).

Real-Time Detection Implementation:

Assess the effort needed to develop a real-time monitoring system for analysing website traffic. Estimate the time required to integrate the trained models into the detection system and implement blocking or warning mechanisms.

System Testing and Validation:

Consider the effort needed for thorough testing of the detection system, including functional and performance testing. Estimate the time required for validation, either through a separate dataset or user feedback.

Deployment and Maintenance:

Assess the effort and resources required for deploying the detection system in the desired environment. Estimate the ongoing effort needed for system maintenance, including regular updates, dataset refreshes, and model retraining.

6. Equations

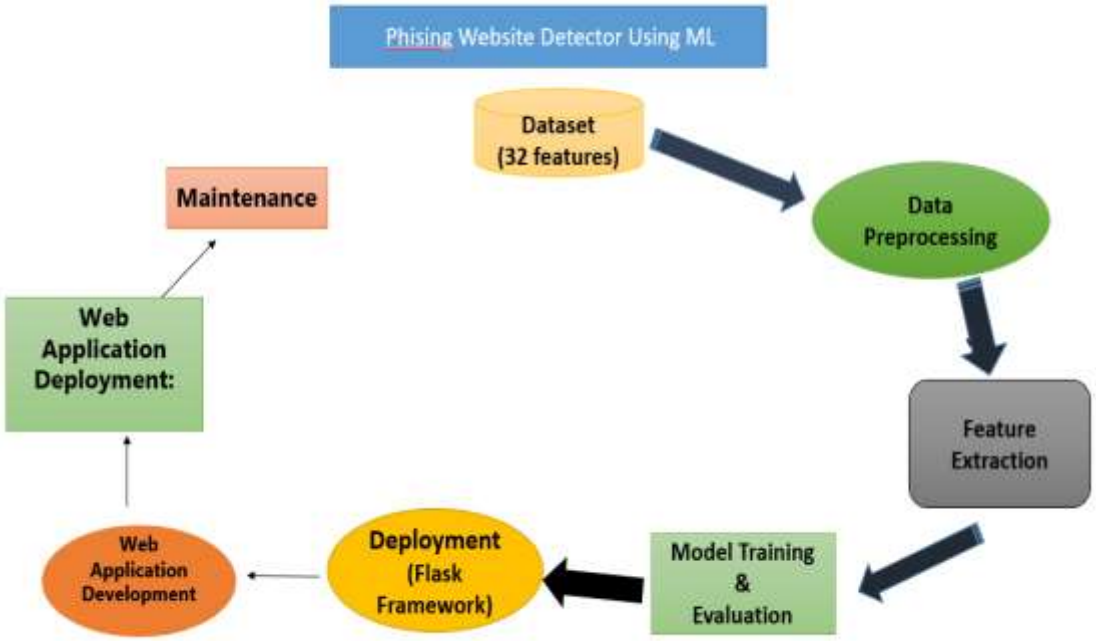

Figure 6.1: Use case diagram

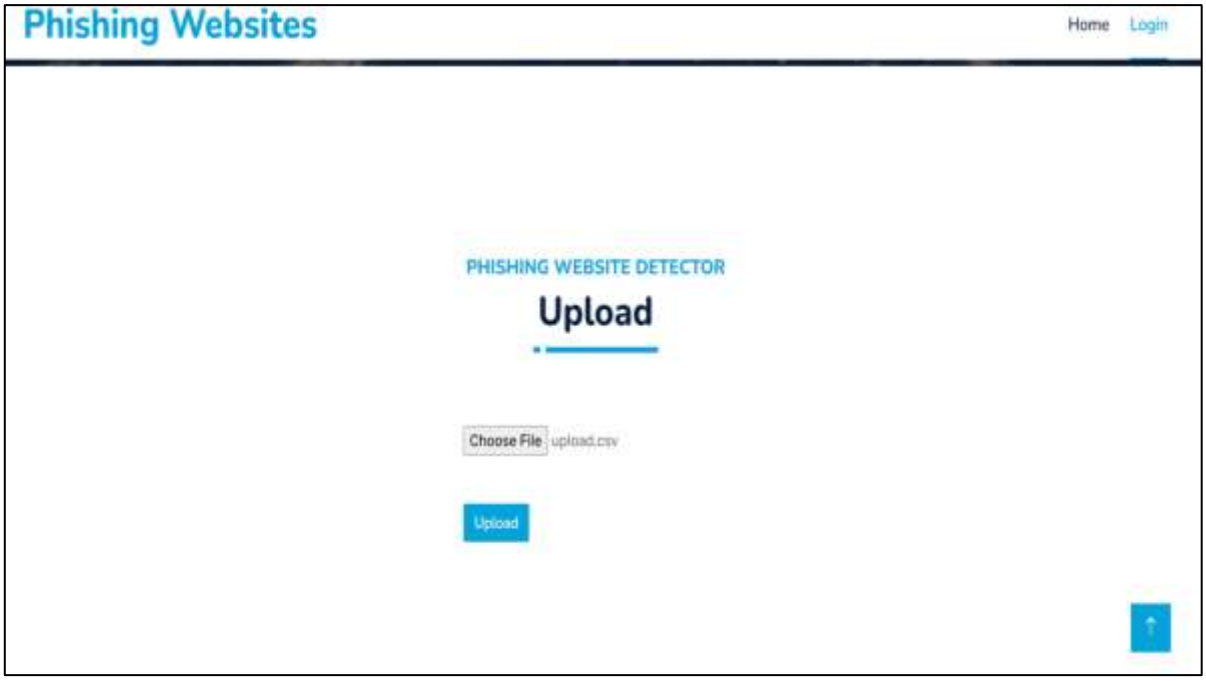

Image 6.2: Dataset Upload

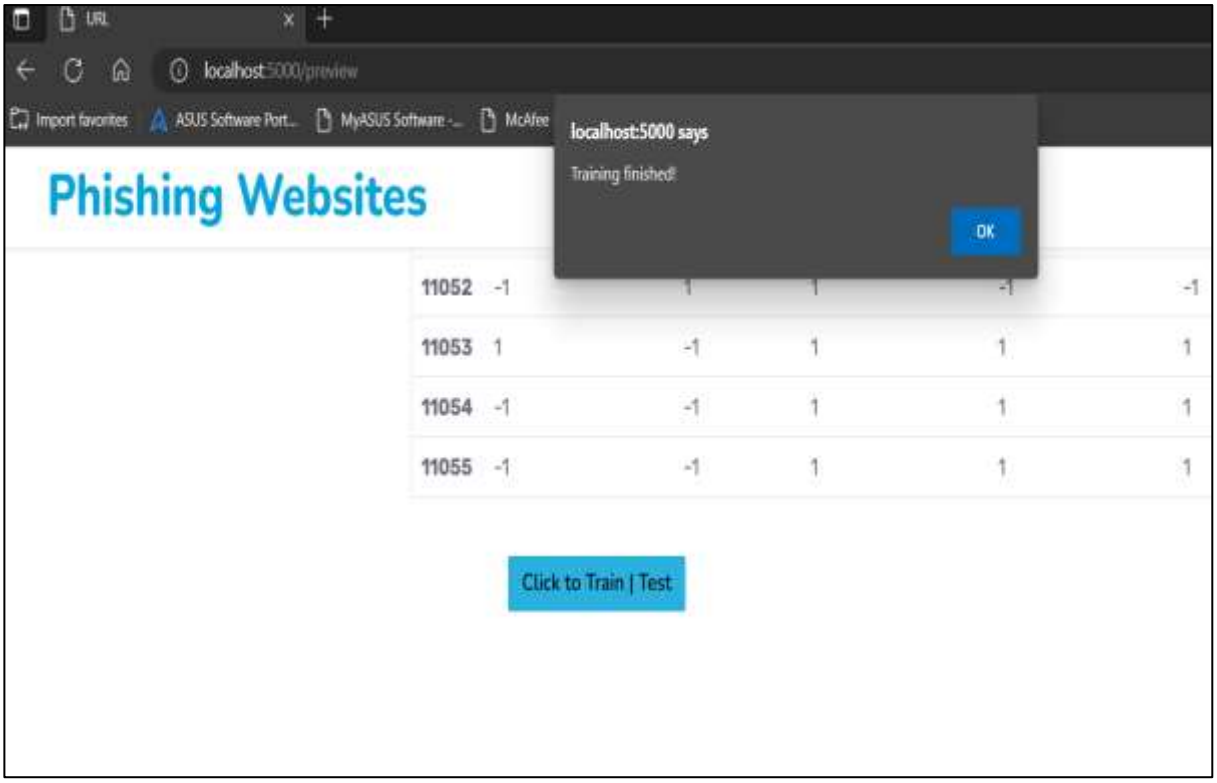

Image 6.3: Dataset Trained

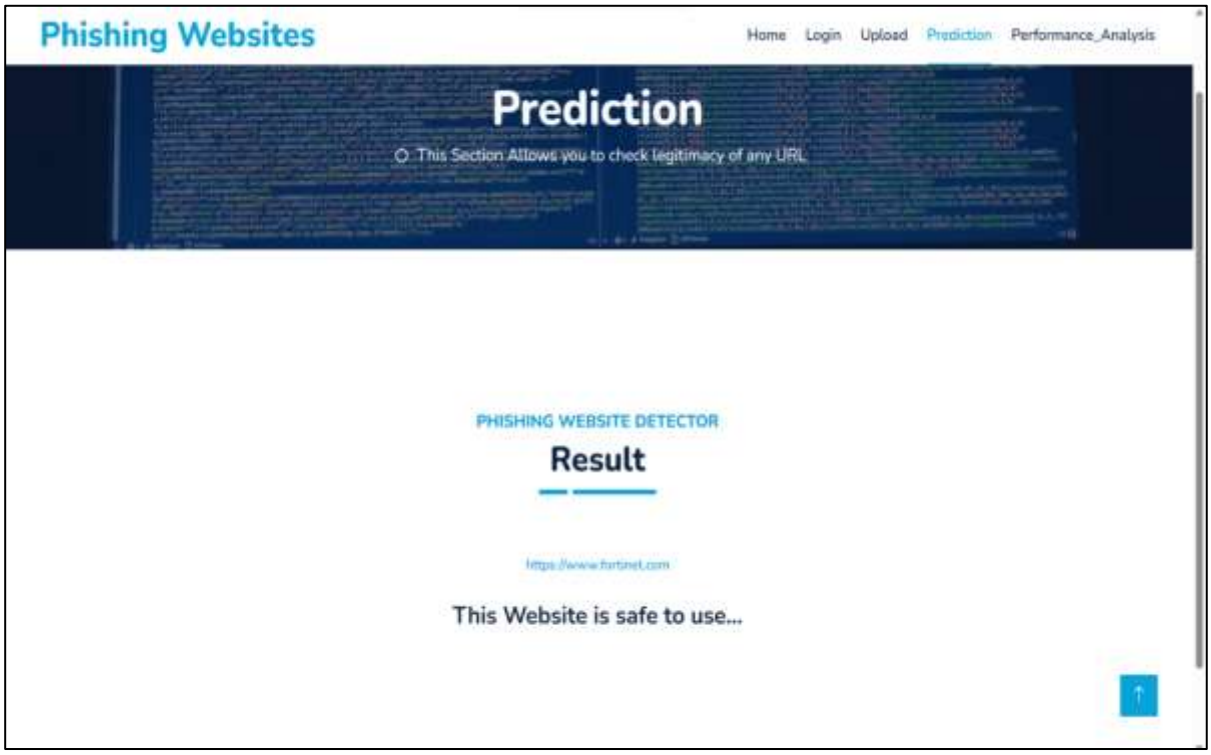

Image 6.4: Prediction Result

7. Conclusion

In recent years, due to the evolving technologies on networking not only for traditional web applications but also for mobile and social networking tools, phishing attacks have become one of the important threats in cyberspace. Although most security attacks target system vulnerabilities, phishing exploits the vulnerabilities of human end-users. Therefore, the main defense form for the companies is informing the employees about this type of attack. However, security managers can get some additional protection mechanisms that can be executed either as a decision support system for the user or as a prevention

mechanism on the servers. In this paper, we aimed to implement a phishing detection system by using some machine learning algorithms specifically Random Forest Algorithm and RNN. The proposed systems are tested with some recent datasets in the literature and reached results are compared with the newest works in the literature. The comparison results show that the proposed systems enhance the efficiency of phishing detection and reach very good accuracy rates. As future works, firstly, it aims to create a new and huge dataset for URL-based Phishing Detection Systems to create a safe, user-friendly environment that can detect illegitimate activities. It is possible to report and block a hacker using a phishing website URL and tracing the location of such anonymous hackers as suggested by Author [10]. Awareness can be created among users by displaying a certain type of Phishing URLs available or causing more harm to our system like zero-hour phishing websites.

References

Jain A.K., Gupta B.B. "PHISH-SAFE: URL Features-Based Phishing Detection System Using Machine Learning", Cyber Security.Advances in Intelligent Systems and Computing, vol. 729, 2018, https://doi.org/10.1007/978-981-10-8536-9_44. Anti-Phishing Working Group (APWG)https://docsapwg.org//reports/apwg_trends_report_q4_2019

Purbay M., Kumar D, "Split Behaviour of Supervised Machine Learning Algorithms for Phishing URL Detection", Lecture Notes in Electrical Engineering, vol. 683, 2021, https://doi.org/10.1007/978-981-15-6840-4_40

Gandotra E., Gupta D, "An Efficient Approach for Phishing Detection using Machine Learning", Algorithms for Intelligent Systems, Springer, Singapore, 2021, https://doi.org/10.1007/978-981-15-8711-5_12

Hung Le, Quang Pham, Doyen Sahoo, and Steven C.H. Hoi, "URL Net: Learning a URL Representation with Deep Learning for Malicious URL Detection",Conference'17, Washington, DC, USA, arXiv:1802.03162 , July 2017.

Hong J., Kim T., Liu J., Park N., Kim SW, "Phishing URL Detection with Lexical Features and Blacklisted Domains", Autonomous Secure Cyber Systems. Springer, https://doi.org/10.1007/978-3-030-33432-1_12.