**Phishing URL Detection using Machine Learning**

**PROJECT SYNOPSIS**

**BACHELOR OF TECHNOLOGY**

**( SEMESTER – VII )**

COMPUTER SCIENCE & ENGINEERING

SUBMITTED BY

**Mr.** **Shridhar Prakash Aware.**

**Ms. Anamika Dhananjay Gulumkar.**

**Ms. Harshada Dattatray Jadhav.**

**Ms. Sakshi Chandashekhar Shinde.**

UNDER THE SUPERVISION OF

**PROF. Mr. Pathak P.A.**



**GAVALI COLLEGE OF ARVIND ENGINEERING SATARA**

**2023 – 2024**

**CERTIFICATE**

This is to certify that the synopsis entitled “**Phishing URL Detection using Machine Learning**” is a bonafide work carried out by,

Ms. Bhagyashri Sunil Yadav.

Ms. Divya Ananda Yadav.

under our supervision during the academic year 2023-24 and submitted to the faculty of Computer Science & Engineering, AGCE, Satara in partial fulfillment of the requirements for the final year of Bachelor of Technology in Computer Science & Engineering.

Mr. Vishwanath Metkari Prof. Pathak P.A

Alumni Mentor Guide

Mr. Gujar V. B

Project Co-Ordinator HOD

Dr. Mirajkar. G. S Dr. Pharande V. A.

Dean R & D Principal

## Table of Contents

|  |
| --- |
|  |
| 1. Introduction |
| 2. Objective and Scope of the Project |
| 3. Literature Review |
| 4. Methodology |
| 5. Hardware Required |
| 6. Software Required |
| 7. Estimation |
| 8. Project Timeline |
| References [IEEE Format] |

**1. Introduction**

## Phishing attacks, a prevalent cybersecurity threat, involve deceiving individuals into disclosing sensitive information by impersonating legitimate websites. As technology advances, so too do the methods employed by malicious actors. To counter this evolving threat landscape, the deployment of machine learning models has emerged as a promising approach. This study introduces a novel framework for the detection of phishing URLs using machine learning techniques.

## The primary objective of this research is to develop a robust and adaptive system that can accurately identify phishing URLs in real-time, providing a proactive defense against cyber threats. The project combines an extensive dataset of both legitimate and malicious URLs with various feature extraction techniques to create a foundation for supervised learning.

## Machine learning algorithms such as decision trees, random forests, support vector machines, and neural networks are trained and fine-tuned to recognize patterns in URLs that distinguish phishing attempts from legitimate websites.

## The proposed system encompasses several key components:

## 1. Data Collection and Preprocessing:

## • A diverse dataset of URLs is collected, containing examples of both phishing and legitimate websites.

## • Data preprocessing techniques are applied to clean and prepare the dataset for machine learning.

## 2. Feature Extraction:

## • Extracting relevant features from the URLs, including domain information, URL length, presence of suspicious keywords, and more.

## 3. Model Selection and Training:

## • Utilizing a variety of machine learning algorithms to create and train models on the dataset.

## • Employing cross-validation to assess model performance and select the most effective algorithms.

## 4. Real-time Detection:

## • Implementing the trained models in a real-time system that can analyze URLs as they are encountered.

## Objective and Scope of the Project

1.Enhance Online Security: The primary objective is to contribute to enhancing online security by proactively detecting phishing attacks. Phishing is a significant cybersecurity threat, and the system's goal is to reduce the risk and impact of such attacks.

2.Real-time Detection: Create a system capable of analyzing URLs as they are encountered, providing immediate protection against phishing attempts for users and organizations.

3.Accuracy and Precision: Develop machine learning models and algorithms that can accurately distinguish between legitimate and phishing URLs, minimizing false positives and false negatives. High accuracy is essential to ensure that genuine websites are not mistakenly flagged as phishing, and that actual phishing sites are correctly identified.

4.Adaptability: Design the system to adapt to evolving phishing tactics. Phishing attacks change over time, so the system should be capable of learning and evolving with new data and attack methods. This adaptability ensures that the system remains effective in the face of changing threats.

5.Efficiency: Build a system that is efficient in processing and classifying URLs, making it practical for real-world use, such as in web browsers, email filters, and network security applications.

6.User and Organizational Protection: Protect users and organizations from falling victim to phishing attacks, which can result in data breaches, financial losses, and reputational damage. The primary aim is to minimize the success rate of phishing attempts.

7.Continuous Learning: Implement mechanisms for continuous learning and improvement. This involves regularly updating and retraining machine learning models with new data to maintain high detection rates.

8.Evaluate Performance: Use rigorous evaluation metrics, such as accuracy, precision, recall, and F1-score, to assess the system's performance and measure its effectiveness in identifying phishing URLs.

9.Reduce Cybersecurity Risks: Ultimately, the overarching objective is to contribute to the reduction of cybersecurity risks associated with phishing attacks, thereby making online activities safer for individuals and organizations.

**3. Literature Review**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sr. No** | **Author** | **Methodology** | **Year** | **Abstract** |
| **1.** | Rishikesh Mahajan  And  Irfan Siddavatam | “The general method to detect phishing websites by updating  blacklisted URLs, Internet Protocol (IP) to the antivirus  database which is also known as “blacklist" method.” | 2018 | 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. Cyber security persons are now looking for  trustworthy and steady detection techniques for phishing  websites detection. This paper deals with machine learning  technology for detection of phishing URLs by extracting and  analyzing various features of legitimate and phishing URLs. |
| **2.** | [P. Amba Bhavani](https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5523208),[Madhumitha Chalamala](https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5440772),[Pinnam Sree Likhitha](https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5523209),[Chanda Pranav Sai Sai](https://papers.ssrn.com/sol3/cf_dev/AbsByAuth.cfm?per_id=5523210) | ss," “Phishing websites are detected using CNN LSTM & CNN Bi-LSTM, Logistic regression and XGBoost algorithms.” | 2022 | 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. As a result, a massive amount of data is constantly downloaded and transferred to the Internet. This allows attackers to access sensitive personal or financial data such as usernames, passwords, account numbers, and social security numbers. |
| **3.** | Rutul Patel, Sanjay Kshetry, Sanket Berad, Justin Zirthantlunga | There are several machine learning algorithms such as Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, and K-Nearest-Neighbor for detecting phishing websites. | 2022 | As we have moved the majority of our monetary, business related, and other day by day exercises to the web, we are presented to more serious dangers as cybercrimes. URL-based phishing assaults are quite possibly the most widely recognized dangers to web client. In this kind of assault, the aggressor takes advantage of the human weakness rather than programming defects. It targets the two people and associations, instigates them to tap on URLs that look secure, and take private data or infuse malware on our framework. |
| **4.** | Mr. Dilip Kumar Jang Bahadur, Sunil D. Kale. | “Method of machine learning model for detecting whether a URL is malicious or not, algorithms such as Random Forests, Decision Trees, Light GBM, Logistic Regression, and Support Vector Machine (SVM) are used.” | 2022 | Phishing is among the most concerning issues in a constantly changing world. The increasing use of the Internet has led to a new way of stealing data, known as [cybercrime](https://www.sciencedirect.com/topics/computer-science/cybercrime). Cybercrime refers to stealing private information and violating privacy through computers. The primary technique used is phishing. Phishing via URLs (Uniform Resource Locators) is one of the most common types, and its primary goal is to steal the data from the user when the user accesses the malicious website. Detecting a malicious URL is a significant challenge. This work aims to provide a solution for detecting such websites with the help of [machine learning algorithms](https://www.sciencedirect.com/topics/computer-science/machine-learning-algorithm) focused on the behaviors and qualities of the suggested URL. |

**4. Methodology**

There are several types of machine learning algorithms that can be used for phishing detection, including supervised learning, unsupervised learning, and deep learning. **Supervised learning algorithms** are trained on labelled data, where the features of each website are used to classify it as either legitimate or phishing. **Unsupervised learning algorithms,** on the other hand, cluster websites based on their features, allowing the detection of outliers that may be indicative of phishing websites.

**Deep learning algorithms, such as convolutional neural networks (CNNs)**, use complex neural network architectures to analyze website features and make predictions.

When training machine learning algorithms for phishing detection, it is important to use a large and diverse dataset of websites. This will help ensure that the algorithms are able to learn and detect phishing websites that are representative of the various types of phishing attacks that exist. Additionally, the features used by the algorithms to distinguish between legitimate and phishing websites must be carefully selected. Common features used in phishing detection include URL structure, website content, and visual cues such as the use of official logos or security certificates.

**5. Hardware Required**

• RAM:- 4 GB

• STORAGE :- 512 GB SSD

• OS :- WINDOWS 7 and above

## 6. Software Required

1. **Development**

* VS code
* Python
* HTML
* CSS

1. **Libraries**

* Flask
* Numpy
* Pandas

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr. No.** | **Particulars** | **Cost** | **Remark** |
| 1 | Operating system | 4000 | (Windows XP,7,8,9,10,11) |
| 2 | Development | 8000 | Includes laptops, software’s etc. |
| 3 | Deployment and Hosting | 10000 | Server and domain name |
| 4 | Total Cost | 22000 | Approximately |

**7. Estimation**

**8. Project Timeline**

Domain & Title of Project

Requirement Analysis and Information Gathering

(Data Collection)

Design, Development and testing

Documentation & Demonstration

## 9.References

**Research Paper :**

[1] Gunter Ollmann, “The Phishing Guide Understanding &

Preventing Phishing Attacks”, IBMInternet Security

Systems, 2007.

[2] https://resources.infosecinstitute.com/category/enterprise

/phishing/the-phishing-landscape/phishing-data-attack-

statistics/#gref

[3] Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A

Literature Survey IEEE, and Andrew Jones, 2013

[4] Mohammad R., Thabtah F. McCluskey L., (2015)

Phishing websites dataset. Available:

https://archive.ics.uci.edu/ml/datasets/Phishing+Websites

Accessed January 2016

[5] http://dataaspirant.com/2017/01/30/how-decision-tree-

algorithm-works/

[6] http://dataaspirant.com/2017/05/22/random-forest-

algorithm-machine-learing/

[7] https://www.kdnuggets.com/2016/07/support-vector-

machines-simple-explanation.html

[8] www.alexa.com

[9] www.phishtank.com