International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Phishing Website Detection Using Machine Learning

Jayesh Pravin Rajput[1], Rajeshwari Yogaraj Chaudhary[2], Shweta Sanjay Avhad[3],Harshada Vilas Patil[4]

***Mr.Liladhar . M. Kuwar.* (*Asst. Prof*)**

Department of Computer Science & Engineering, D.N.Patel College Of Engineering, Shahada

Jayesh Pravin Rajput, Computer Science & Engineering, D.N.Patel College Of Engineering, Shahada, Rajeshwari Yogaraj Chaudhary, Computer Science & Engineering, D.N.Patel College Of Engineering, Shahada, Shweta Sanjay Avhad, Computer Science & Engineering, D.N.Patel College Of Engineering, Shahada, Harshada Vilas Patil, Computer Science & Engineering, D.N.Patel College Of Engineering, Shahada.

Abstract – Phishing is a deceptive technique used by cybercriminals to impersonate legitimate websites and obtain sensitive user information such as login credentials and financial data. This project proposes a machine learning-based approach to detect phishing websites by analyzing URL features and content-related indicators. Various algorithms, including Logistic Regression, Multinomial Naïve Bayes, and XGBoost, were evaluated for their classification accuracy. The system employs preprocessing techniques such as tokenization, stemming, and vectorization using Count Vectorizer to convert textual URL data into machine-readable formats. FastAPI is utilized for seamless deployment of the trained model, providing a user-friendly interface. Among the tested models, Logistic Regression achieved the highest accuracy of 96.63%, demonstrating its effectiveness in phishing detection. The proposed solution not only enhances detection accuracy but also enables real-time identification of phishing threats, contributing to a safer online environment.

Key Words: Phishing Detection, Logistic Regression, XGBoost, Multinomial Naïve Bayes, URL Features, FastAPI, Machine Learning.

1.INTRODUCTION

The rapid expansion of the internet has significantly enhanced global communication and digital transactions. However, this growth has also led to a rise in cyber threats, with phishing attacks emerging as one of the most prevalent and damaging forms. Phishing is a technique used by attackers to deceive users into revealing sensitive information—such as usernames, passwords, and banking credentials—by mimicking legitimate websites or organizations through fake emails and websites. These fraudulent pages often replicate real brands using official logos, layouts, and domain names to gain user trust.

The economic and security impact of phishing is substantial, with billions of dollars lost annually due to compromised user accounts and financial fraud. Traditional security mechanisms and blacklists are often insufficient, as phishers continually adapt their tactics to bypass static detection methods.This project focuses on the development of a machine learning-based system to detect phishing websites effectively. By leveraging supervised learning algorithms—such as Logistic Regression, Multinomial Naïve Bayes, and XGBoost—the system analyzes URL features, host data, and page content characteristics to classify websites as either legitimate or malicious. The objective is to build a robust and scalable detection model with high accuracy and low latency.

Furthermore, the system incorporates a user-friendly web interface built using FastAPI, enabling real-time predictions based on user-input URLs. The project not only demonstrates the practical application of machine learning in cybersecurity but also addresses current challenges in phishing detection by comparing multiple models and selecting the best-performing approach.

2. LITERETURE SURVEY

A.Phishing Website Detection using Machine Learning Algorithms

Authors : Mahajan, Rishikesh & Siddavatam, Irfan. (2018). Phishing Website Detection using Machine Learning Algorithms. International Journal of Computer Applications.

Summary : This Paper Shows 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.Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites.

B. PHISHING WEBSITE DETECTION

Authors : Akriti Soni, Pranchal Abrol, Department of Computer Science & Engineering and Information Technology Jaypee University of Information Technology, Waknaghat, 173234, Himachal Pradesh, INDIA

Summary : This paper shows how Online phishing is one of the most common attacks on the modern internet. The goal of phishing website uniform resource locators is to steal personal data including login credentials and credit card numbers. As technology keeps growing, phishing strategies began to develop rapidly.

C. A systematic literature review on phishing website detection techniques

Authors : Asadullah Saf, Satwinder Singh, Dept. of Computer Science & Technology, Central University of Punjab, Bathinda, Punjab, India, Nangarhar University, Ministry of Higher Education, Afghanistan

Summary : In this paper we see that Phishing is a fraud attempt in which an attacker acts as a trusted person or entity to obtain sensitive information from an internet user. In this Systematic Literature Survey (SLR), different phishing detection approaches, namely Lists Based, Visual Similarity, Heuristic, Machine Learning, and Deep Learning based techniques, are studied and compared. For this purpose, several algorithms, data sets, and techniques for phishing website detection are revealed with the proposed research questions

D. Phishing Website Detection Using Deep Learning Models

Authors : UME ZARA, KASHIF AYYUB, HIKMAT ULLAH KHAN, ALI DAUD, TARIQ ALSAHFI , AND SAIMA GULZAR AHMAD, Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah 23218, Saudi Arabia

Summary : In this paper we see that research addresses the imperative need for advanced detection mechanisms for the identification of phishing websites. For this purpose, we explore state-of-the-art machine learning, ensemble learning, and deep learning algorithms. Cybersecurity is essential for protecting data and networks from threats. Detecting phishing websites helps prevent fraud and safeguard personal information. To evaluate the efficacy of our proposed method, the top features using information gain, gain ratio, and PCA are used to predict and identify a website as phishing or non-phishing. The proposed system is trained using a dataset that covers 11,055 websites. The ensemble learning model applied achieved an impressive 99% accuracy in predicting phishing websites, surpassing previous models, and setting a new benchmark in the field.

E. Phishing or Not Phishing? A Survey on the Detection of Phishing Websites

Authors : R. Zieni, L. Massari and M. C. Calzarossa, "Phishing or Not Phishing? A Survey on the Detection of Phishing Websites," in *IEEE Access*, vol. 11, pp. 18499-18519, 2023, doi: 10.1109/ACCESS.2023.3247135.

Summary : In this paper we see that Phishing is a security threat with serious effects on individuals as well as on the targeted brands. Although this threat has been around for quite a long time, it is still very active and successful. In fact, the tactics used by attackers have been evolving continuously in the years to make the attacks more convincing and effective. In this context, phishing detection is of primary importance. The literature offers many diverse solutions that cope with this issue and in particular with the detection of phishing websites.

F. Cyber-Phishing Website Detection Using Fuzzy Rule Interpolation

Authors : Mohammad Almseidin, Mouhammad Alkasassbeh, Maen Alzubi, Jamil Al-Sawwa, Department of Computer Science, Princess Sumaya University for Technology, Amman 11941, Jordan

Summary : This paper introduces a novel detection method for phishing website attacks while avoiding the issues associated with the deficiencies of the knowledge-based representation and the binary decision. The suggested detection method was performed using Fuzzy Rule Interpolation (FRI). The FRI reasoning methods added the benefit of enhancing the robustness of fuzzy systems and effectively reducing the system's complexity. These benefits help the Intrusion Detection System (IDS) to generate more realistic and comprehensive alerts in case of phishing attacks. The proposed method was applied to an open-source benchmark phishing website dataset. The results show that the proposed detection method obtained a 97.58% detection rate and effectively reduced the false alerts. Moreover, it effectively smooths the boundary

between normal and phishing attack traffic because of its fuzzy nature.

3.OBJECTIVES

- **To develop a machine learning-based system** capable of accurately detecting phishing websites by analyzing URL and webpage features.
- **To compare the performance of multiple classification algorithms** including Logistic Regression, Multinomial Naïve Bayes, and XGBoost in order to identify the most effective model for phishing detection.
- **To extract and preprocess relevant URL features** such as tokenized words, suspicious characters, and domain patterns that help in distinguishing between legitimate and phishing websites.
- **To implement a user-friendly interface** using FastAPI for real-time input and detection, allowing users to easily verify the authenticity of any given URL.
- **To improve detection accuracy and reduce false positives**, ensuring that the selected model is both reliable and efficient in practical cybersecurity applications.
- **To deploy the final trained model** in a way that supports quick response times, high scalability, and ease of integration into broader security systems.

4.METHODOLOGY

The methodology for developing the phishing website detection system is structured into several key stages, beginning with data preprocessing. Initially, a dataset containing labeled URLs (phishing or legitimate) is collected and prepared for analysis. Preprocessing involves tokenizing the URLs using regular expressions to break them into meaningful components, followed by stemming to reduce words to their root forms, making the data more consistent. These processed text features are then converted into numerical vectors using Count Vectorizer, enabling the machine learning models to interpret and learn from the data.

Next, feature extraction is performed to identify and highlight characteristics that are commonly associated with phishing URLs. This includes examining suspicious keywords (like "secure" or "login"), the presence of special characters, URL length, number of dots, and domain-based attributes. These features are essential in distinguishing between benign and malicious websites.

Following this, three machine learning algorithms—Logistic Regression, Multinomial Naïve Bayes, and XGBoost—are trained and tested using the extracted features. Each model's performance is evaluated based on its accuracy, with Logistic Regression emerging as the best-performing model, achieving an accuracy rate of 96.63%.

To make the system usable in real-world scenarios, the trained model is deployed using FastAPI, a modern and high-performance Python framework. This allows users to input a URL via a web interface and receive instant predictions about its legitimacy. Prior to deployment, the entire system undergoes rigorous testing to ensure accuracy, reliability, and compatibility with hardware and software requirements. The methodology ensures that the system is both efficient and user-friendly while maintaining high detection performance.

CONCLUSION

Phishing remains a significant threat to internet users, leading to the theft of sensitive personal and financial information. The project effectively addresses this challenge by developing a machine learning-based system that detects phishing websites using key URL features and content attributes. By comparing various classification models—Logistic Regression, Multinomial Naïve Bayes, and XGBoost—the system identifies Logistic Regression as the most accurate model, achieving an impressive accuracy of 96.63%. The implementation of advanced preprocessing techniques such as tokenization, stemming, and vectorization further enhances the model's performance.

In addition to accurate detection, the system is designed with practical usability in mind. The integration of FastAPI enables a responsive and accessible web interface, allowing users to input URLs and receive real-time classification results. Compared to traditional approaches, this solution offers improved accuracy, lower latency, and a user-friendly interface.

Overall, the project demonstrates that machine learning is a powerful tool for combating phishing attacks. With continuous updates and integration into cybersecurity infrastructures, such systems can play a vital role in protecting users from online fraud and identity theft.

REFRENCES

I. https://www.researchgate.net/publication/328541785_Phishing_Website_Detection_using_Machine_Learning_Algorithms
II. http://www.ir.juit.ac.in:8080/jspui/bitstream/123456789/3820/1/Phishing%20Website%20Detection.pdf
III. https://www.sciencedirect.com/science/article/pii/S1319157823000034
IV. https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10735206
V. https://ieeexplore.ieee.org/document/10049452
VI. https://www.mdpi.com/2410-387X/6/2/24