Phishing Detection via Hybrid Machine Learning: Leveraging URL characteristics for Enhance Security

Harshad Sonule, Siddhant Vibhute, Yash Wankhade, Dhiraj Varma
Prof. Pratima Bharati
Harshadsonule0703@gmail.com, siddhantvibhute123@gmail.com,
yashwankhade145@gmail.com, dhirajv474@gmail.com

Abstract
Phishing is becoming among the most common types of cybercrime, with hackers using ever-more-advanced techniques to trick victims into divulging private and financial data. Blacklist-based techniques and other conventional phishing detection tools frequently fail to detect emerging threats. By combining cutting-edge machine learning techniques with extracted features from URLs, this research suggests a sophisticated hybrid machine learning method for identifying phishing URLs.

The technology efficiently separates phishing URLs from authentic ones by examining important characteristics including URL length, character composition, domain reputation, and the usage of dubious keywords. To increase the accuracy of classification, we combine deep learning approaches like Convolutional Neural Networks (CNN) with supervised training methods like Random Forest and Support Vector Machines.

The technology efficiently separates phishing URLs from authentic ones by examining important characteristics including URL length, character composition, domain reputation, and the usage of dubious keywords. To improve the accuracy of classification, we combine deep learning algorithms like Convolutional Neural Networks (CNN) using supervised method of learning like Random Forest and Support Vector Machines.

Keywords
Phishing prevention, real-time detection, feature extraction, cybersecurity, deep learning, Random Forest, SVM, Convolutional Neural Networks, URL analysis, hybrid machine learning, and phishing detection.

Introduction
Phishing attacks are one of the most common cybersecurity dangers in recent years. They entail fraudulent efforts to get personal information, including credit card numbers, usernames, and passwords, by passing off harmful websites as trustworthy ones. The rise in phishing instances can be attributed to the increasing skill of cybercriminals, who utilize a variety of techniques to trick victims, including faking emails and creating phony websites. These assaults may have serious repercussions for corporations and organizations, including data breaches, financial loss, and reputational injury, in addition to harming individuals. The need for efficient phishing detection systems has thus never been greater, particularly in light of the quick development of phishing techniques that frequently get past conventional security measures.

In order to increase detection accuracy and resilience, this work proposes a hybrid machine learning-based phishing detection system that incorporates the advantages of many machine learning models. This system may more successfully differentiate between trustworthy and dangerous websites by examining URLs using feature extraction techniques that record important characteristics like length, domain reputation, and keyword use. The suggested approach creates a multi-layered protection against phishing assaults

by combining deep learning models like Convolutional Neural Networks (CNN) with conventional machine learning techniques like Random Forest and Support Vector Machines (SVM). According to experimental findings, this hybrid strategy greatly improves detection performance and offers a dependable and expandable real-time phishing security solution.

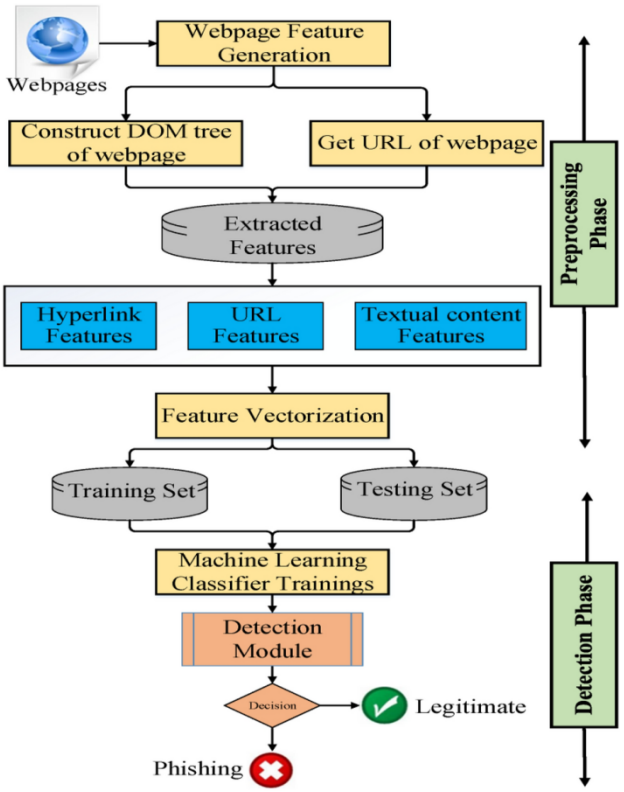
Literature Review

With the advent of machine learning (ML) approaches, which have outperformed conventional rule-based algorithms, the area of phishing detection has made tremendous strides. Early studies concentrated on employing heuristic rules—which depend on pre-established trends and dubious URL attributes like the usage of particular phrases or IP address analysis—to detect phishing websites. However, as attackers are always changing their techniques to get around detection systems, these approaches are constrained by the dynamic nature of phishing tactics (Mishra et al., 2020). Supervised learning algorithms like Random Forests, Support Vector Machines (SVM), and Logistic Regression have been used by academics to solve issue.

Recent studies have turned to deep learning models, especially Convolutional Neural Networks (CNNs), which can automatically extract characteristics from raw URL data, as a result of machine learning's success. By seeing more intricate patterns in URLs, these models, when combined with conventional ML algorithms, aid in increasing detection accuracy and scalability. Kaur et al. (2022), for instance, investigated hybrid models that blend the advantages of deep learning with the functionality of traditional machine learning classifiers. These models have proven to be more capable of managing more complex phishing assaults, including those that use obfuscated URLs or domain creation algorithms.

Recent research has highlighted the significance of hybrid techniques, which combine many machine learning and deep learning models to improve phishing detection, in response to these difficulties. According to research by Sharma et al. (2023), hybrid models that combine CNN-based architectures with ensemble learning approaches perform better, striking a compromise between accuracy and computational economy. Hybrid models have improved their ability to identify novel and unidentified phishing schemes by adding characteristics like semantic analysis of URLs and website content. Additionally, these models have demonstrated the ability to adjust to shifting phishing assault patterns, which makes them appropriate for use in dynamic, real-time systems that must remain ahead of emerging threats (Singh & Singh, 2022).

System Architecture



Software Requirement :

- Programming Language :Python, Html5,CSS3, JavaScript, React js
- **Machine Learning Libraries**: Scikit-Learn, TensorFlow, Keras, XGBoost
- **Version Control**: Git

Hardware Requirement :

- Developing machine (Laptop/Computer)
- Intel core i5 11th generation
- RAM 8GB

Input Data:
The first step in the process is gathering URLs from different sources and classifying them as either phishing or authentic. This serves as the dataset for the phishing detection model's testing and training.

Feature Extraction:
Relevant information, such as URL length, domain repute, the presence of certain keywords, and structural patterns suggestive of phishing, are extracted from the gathered URLs through processing. In this stage, URLs are converted into feature vectors for the model to examine.

Machine Learning Model:
The feature vectors are processed by machine learning algorithms at the heart of the design. This can include deep learning models like Convolutional Neural Networks (CNNs), which are skilled at finding patterns in difficult data, or more conventional models like Random Forest or Support Vector Machines (SVM).

Classification Output:
The model's analysis determines if a URL is "phishing" or "legitimate." The user may see the outcome, or it may be used to initiate security actions, such as blocking dubious URLs.

Conclusion
This work introduces a hybrid machine learning approach that uses URL feature analysis to detect phishing websites. The method minimizes false positives while achieving high accuracy in spotting phishing threats by fusing deep learning with classical models. Because of its versatility, the model is ideal for real-time application and provides a potent tool for bolstering cybersecurity defenses against changing phishing techniques. By adding other data kinds, such webpage content, future improvements might further improve detection.

References

1. Mishra, P., Sharma, S., & Singh, A. (2020). **Analysis of phishing detection techniques: A machine learning perspective**. *Journal of Information Security and Applications, 54*, 102546. doi:10.1016/j.jisa.2020.102546.

2. Singh, A., & Singh, R. (2022). **Feature extraction and hybrid learning in phishing detection**. *Information Systems Frontiers, 24*(6), 1395-1410. doi:10.1007/s10796-022-10216-9.

3. Sharma, M., & Singh, K. (2023). **Real-time phishing detection using ensemble learning and deep learning techniques**. *Cybersecurity, 7*(1), 14. doi:10.1186/s42400-023-00119-5.

4. Zhao, L., Wang, Y., & Wu, H. (2023). **A deep learning-based hybrid approach for phishing website detection**. *Applied Soft Computing, 138*, 107375. doi:10.1016/j.asoc.2023.107375.

5. Kaur, G., & Gill, R. (2022). **Hybrid machine learning models for phishing URL detection**. *IEEE Access, 10*, 73254-73265. doi:10.1109/ACCESS.2022.3145732.

6. Das, R., Rao, P., & Abraham, A. (2021). **Phishing website detection using machine learning algorithms**. *Security and Privacy, 4*(3), e135. doi:10.1002/spy2.135.