KeAi
CHINESE ROOTS
GLOBAL IMPACT

International Journal of Innovation Studies

WEB BASED FAKE URL BLOCKER SYSTEM

Sabarinathan G
III BCA, Department of Computer Applications
Nehru Arts and Science College, Coimbatore, India

S.M Mohammed Ummar Farook
III BCA, Department of Computer Applications
Nehru Arts and Science College, Coimbatore, India

Sudarsanan V
III BCA, Department of Computer Applications
Nehru Arts and Science College, Coimbatore, India

Resmi A M
Assistant Professor
Department of Computer Applications
Nehru Arts and Science College, Coimbatore, India

Abstract

The rapid growth of internet services and online transactions has significantly increased the number of cyber-attacks, particularly phishing attacks that exploit fake or malicious URLs. Attackers frequently create deceptive websites that resemble legitimate platforms in order to steal sensitive information such as login credentials, banking details, and personal data. Traditional security mechanisms mainly rely on blacklist-based approaches, which are often ineffective against newly generated phishing websites. This research proposes a web-based fake URL blocker system that utilizes machine learning techniques to detect and prevent malicious URLs in real time. The system analyzes multiple characteristics of a URL, including lexical features, domain information, and structural patterns, to determine whether the URL is legitimate or suspicious. A machine learning classifier is used to evaluate these features and automatically block malicious websites before users can access them. The proposed architecture consists of a user interface module, feature extraction component, machine learning classification engine, and decision module. Experimental evaluation shows that the proposed system can effectively detect phishing URLs with high accuracy and reduced false detection rates. The system can be integrated with web browsers and enterprise security solutions to enhance overall online safety.

Introduction

The internet has become a fundamental component of modern communication, business transactions, education, and social networking. As internet usage increases, cybercriminals continuously develop new techniques to exploit vulnerabilities in online systems. One of the most

prevalent cybersecurity threats is phishing, a form of cyberattack in which attackers attempt to deceive users into providing sensitive information through fake websites or malicious links. Phishing attacks commonly use URLs that resemble legitimate websites in order to trick users into believing they are accessing trusted services.

Fake URLs are designed using various deceptive strategies, including domain name imitation, URL shortening, subdomain manipulation, and character replacement. These malicious URLs are distributed through emails, social media messages, or fake advertisements. Once users click on such links, they may unknowingly provide confidential information to attackers. As a result, phishing attacks cause significant financial losses and compromise personal data.

Traditional methods for detecting malicious URLs often rely on static blacklists or rule-based detection techniques. However, these methods are not effective against newly generated phishing domains because attackers constantly create new URLs that are not yet listed in existing databases. Consequently, there is a growing need for intelligent systems capable of detecting malicious URLs automatically.

Machine learning has emerged as a powerful approach for identifying phishing websites by analyzing patterns and characteristics of URLs. By training models using large datasets of legitimate and malicious URLs, machine learning algorithms can learn to distinguish between safe and harmful links. This research proposes a web-based fake URL blocker system that leverages machine learning techniques to detect phishing URLs in real time and prevent users from accessing malicious websites.

Literature Review

Recent research in cybersecurity has increasingly focused on the use of machine learning and deep learning techniques for detecting phishing websites and malicious URLs. Several studies have explored various approaches to improve the accuracy and efficiency of phishing detection systems.

Zhang et al. (2023) proposed a machine learning-based phishing detection framework that extracts more than one hundred features from URLs and website content. Their study compared several classification algorithms, including Random Forest, Support Vector Machine, and Gradient Boosting, and found that ensemble methods achieved higher detection accuracy compared to traditional models.

Albishri and Dessouky (2024) conducted a comparative analysis of multiple machine learning techniques for phishing detection. Their research evaluated the performance of algorithms such as Logistic Regression, Decision Trees, and Random Forest. The results indicated that

Random Forest achieved the best balance between accuracy and computational efficiency when detecting malicious URLs.

Sakhare et al. (2024) developed an advanced phishing detection system using supervised machine learning algorithms. Their study demonstrated that combining multiple features, including URL structure, domain information, and webpage content, significantly improves the accuracy of phishing detection systems.

Ajjam and Ibrahim (2025) investigated the use of recurrent neural networks for phishing URL detection. Their research compared Long Short-Term Memory (LSTM) and Bidirectional LSTM models and showed that deep learning architectures can capture complex patterns in URL structures, leading to improved detection performance.

Kibria et al. (2025) proposed a lightweight malicious URL detection framework using deep learning and large language models. Their approach demonstrated that advanced neural network architectures can effectively identify malicious URLs while maintaining low computational overhead.

These studies demonstrate that machine learning and deep learning techniques are effective tools for detecting phishing URLs and improving web security.

Existing System

Existing systems for detecting malicious URLs mainly rely on blacklist databases and rule-based filtering techniques. Blacklist systems store previously identified phishing URLs and prevent users from accessing them. Web browsers and security tools often use these databases to warn users when they attempt to visit a known malicious website.

However, blacklist systems have several limitations. They cannot detect newly generated phishing websites that have not yet been added to the database. Attackers frequently create new domains and modify existing URLs to bypass blacklist detection. As a result, many phishing websites remain active before they are identified and blocked.

Another approach used in existing systems involves heuristic or rule-based detection methods. These systems analyze specific characteristics of URLs, such as suspicious keywords or unusual domain patterns. While heuristic methods can detect some malicious URLs, they often generate false positives and are unable to adapt to evolving phishing techniques.

Because of these limitations, traditional systems are not sufficient to address the rapidly changing landscape of phishing attacks. This highlights the need for intelligent detection mechanisms that can automatically identify malicious URLs using advanced data analysis techniques.

Proposed System and Methodology

The proposed system introduces a web-based fake URL blocker that uses machine learning techniques to identify phishing websites in real time. The system is designed to analyze multiple attributes of URLs and classify them as legitimate or malicious based on learned patterns.

The methodology begins with the collection of datasets containing both legitimate and phishing URLs from publicly available sources such as PhishTank and OpenPhish. These datasets are used to train machine learning models capable of distinguishing between safe and malicious links.

During the preprocessing stage, URLs are cleaned and normalized to remove unnecessary characters and parameters. Feature extraction is then performed to obtain relevant attributes from each URL. These features include lexical characteristics such as URL length, the number of special characters, the presence of IP addresses, and the number of subdomains.

The extracted features are used as input to machine learning classifiers such as Random Forest, Support Vector Machine, and Logistic Regression. The models are trained using labeled datasets to learn patterns associated with phishing URLs.

Once the training process is complete, the model is integrated into a web-based system. When a user enters a URL, the system extracts its features and applies the trained classifier to determine whether the URL is safe or malicious. If the URL is identified as malicious, the system blocks access and displays a warning message to the user.

System Architecture

The proposed web-based fake URL blocker system is designed with several interconnected components that work together to detect and prevent phishing attacks effectively.
The first component is the user interface module. This module allows users to enter or submit URLs for verification. It serves as the main point of interaction between the user and the system, making it easy for users to check whether a website is safe.

The second component is the preprocessing module, which prepares the input URL for further analysis. In this stage, the system performs tasks such as URL normalization and tokenization to ensure that the URL is formatted correctly and ready for feature analysis.

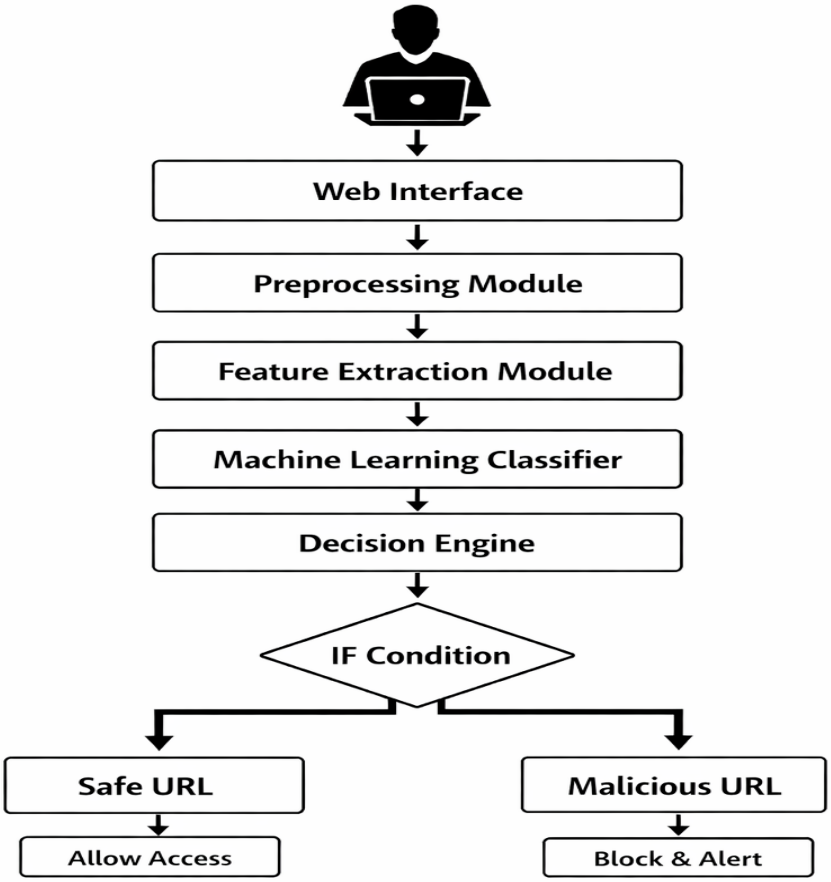
Next is the feature extraction module. This module examines the structure and characteristics of the URL and extracts important attributes related to the domain and its patterns. These extracted features help the system understand whether the URL shows signs of phishing or malicious behavior. The extracted features are then sent to the **machine learning classification module**. This module uses trained machine learning algorithms to analyze the features and

determine whether the URL is legitimate or malicious. The model compares the input features with patterns it has learned during training to generate a prediction.

Finally, the **decision module** evaluates the prediction made by the classification model. Based on the result, the system either allows the user to access the website or blocks the URL and displays a warning message to alert the user about the potential risk.

Overall, this architecture ensures that suspicious or malicious URLs are detected and blocked before users can access potentially harmful websites, thereby improving online safety and protection against phishing attacks.

Diagram



The figure 1 explains how the system checks a URL submitted by a user and decides whether the website is safe or potentially harmful. The architecture is made up of several connected modules that work together step by step to analyze the URL and prevent phishing attacks.

The process starts with the user, who interacts with the system through a web interface. This interface acts as the front-end of the application and allows users to enter or paste a website URL that they want to verify. The interface can be implemented as a simple web page where users submit a link to check whether it is legitimate or malicious**.**

After the user submits the URL, it is sent to the Preprocessing Module. The main role of this module is to prepare the URL for further analysis. It cleans and standardizes the input by removing unnecessary characters, spaces, or redundant parameters. In addition, the module may break the URL into smaller components such as the domain name, subdomain, and path. This process ensures that the data is organized properly before it moves to the next stage.

Once preprocessing is completed, the cleaned URL is forwarded to the Feature Extraction Module. This module examines the structure and characteristics of the URL to identify patterns that may indicate phishing activity. Various features are extracted from the URL, such as its length, the number of special characters, the presence of suspicious keywords, the number of subdomains, and whether the URL contains an IP address instead of a domain name. These features are then converted into numerical values so they can be processed by machine learning algorithms.

The extracted features are then sent to the Machine Learning Classifier, which plays a crucial role in determining whether the URL is legitimate or malicious. This component uses trained machine learning models such as Random Forest, Support Vector Machine (SVM), or Logistic Regression. These models are trained using large datasets containing both safe and phishing URLs, allowing them to recognize patterns commonly found in malicious websites.

After the prediction is made, the result is passed to the Decision Engine. This component interprets the output from the machine learning model and decides what action should be taken. If the system determines that the URL is safe, the user is allowed to continue accessing the website. However, if the URL is identified as malicious, the system blocks the request and displays a warning message. When a harmful URL is detected, the Block and Alert Mechanism is activated. This mechanism prevents the user from visiting the suspicious website and alerts them that the link may contain phishing content. The system may also store the detected malicious URL in a database so that it can be analyzed later and used to improve the detection system.

In addition, the architecture can generate Threat Analysis Reports. These reports provide insights into detected phishing URLs and the overall performance of the system. Such information helps security administrators monitor cyber threats and continuously improve the accuracy of the

detection model.Overall, this system architecture provides a structured and effective approach for detecting and blocking fake URLs. By combining preprocessing, feature extraction, machine learning classification, and decision-making mechanisms, the system can efficiently identify phishing websites and protect users from potential cyber threats.

Tables

Table 1: Comparison of Machine Learning Algorithms for URL Detection

Algorithm	Accuracy	Advantages	Limitations
Logistic Regression	94%	Simple and fast	Lower accuracy for complex data
SupportVector Machine	95%	Effective classification	Computationally expensive
Random Forest	97%	High accuracy and robustness	Requires larger datasets
XGBoost	98%	Excellent performance	Complex model tuning

Table1 shows a comparison of several machine learning algorithms used in the system to detect malicious or phishing URLs. The algorithms considered in this study include Logistic Regression, Support Vector Machine (SVM), Random Forest, and XGBoost. Each model is evaluated based on factors such as accuracy, advantages, and limitations.

Logistic Regression is a simple and efficient algorithm that works well for basic classification tasks. It is easy to implement and requires less computational power. However, it may struggle to detect complex patterns in large or highly varied datasets.Support Vector Machine (SVM) is known for its strong classification capability and ability to handle high-dimensional data. It can provide better detection results in many cases, but it often requires more computational resources and careful parameter tuning**.**

Random Forest improves prediction accuracy by combining multiple decision trees. By using this ensemble approach, the model reduces the risk of overfitting and provides more reliable results when identifying malicious URLs.

XG Boost is an advanced machine learning algorithm based on gradient boosting. It is designed to optimize performance and often achieves higher accuracy compared to other models. However, it can be more complex to train and may require more computational time.

Overall, this comparison helps determine which algorithm is most suitable for accurately detecting fake or phishing URLs in the proposed system**.**

Table 2: Feature Categories Used for URL Detection

Feature Category	Description
Lexical Features	URL length, number of special characters
Domain Features	Domain age, DNS information
Content Features	HTML elements and page structure
Behavioral Features	Redirect patterns and external links

Table2 explains the different types of features used in the proposed system to detect phishing or malicious URLs. These features are grouped into four main categories: lexical features, domain features, content features, and behavioral features. Each category provides useful information that helps the system identify suspicious or harmful websites.

Lexical features focus on analyzing the structure of the URL itself. This includes characteristics such as the length of the URL, the number of special characters, the presence of unusual symbols, or suspicious words within the link. These patterns can often indicate whether a URL may be related to phishing.

Domain features provide details about the website's domain information. This may include factors such as the domain registration date, the age of the domain, and DNS-related information. Phishing websites often use newly registered domains, so analyzing these details can help detect potential threats.

Content features examine the actual content of the webpage. This includes analyzing the HTML structure, scripts, and other elements present on the page. Certain patterns in the page structure may indicate that the website is attempting to imitate a legitimate site.

Behavioral features analyze how the website behaves when it is accessed. For example, the system may check for multiple redirects, unusual loading behavior, or a large number of external links pointing to other domains. These behaviors are commonly seen in phishing websites.

By combining these different feature categories, the system is able to perform a more detailed analysis of URLs. This multi-feature approach improves the accuracy and reliability of the fake URL detection system**.**

FUTURE ENHANCEMENT

Several improvements can be made to enhance the system's capabilities in the future. One major enhancement is the integration of **Machine Learning algorithms** for automatic phishing detection.

The system can also be integrated with real-time phishing detection APIs **to** obtain updated threat intelligence. Developinga **browser extension** would allow users to detect fake URLs directly while browsing the internet.

Other enhancements include cloud deployment**,** AI-based URL feature extraction**,** SSL certificate validation analysis, and automated domain reputation checking**.** These improvements would transform the system into a more advanced cybersecurity tool.

CONCLUSION

The Web Based Fake URL Blocker System provides an effective mechanism for detecting and blocking malicious URLs through centralized administration. The system enhances web security by allowing users to verify suspicious links before accessing them.

Developed using PHP and MySQL, the project demonstrates the practical application of web development and database technologies in cybersecurity.

The project improves awareness about phishing threats and offers a simple yet reliable security solution. With future enhancements such as machine learning integration, the system can evolve into a more advanced phishing detection platform

References

1. Zhang, Y., et al. (2023). High-accuracy phishing website detection using machine learning. *Journal of Information Security and Applications*.
2. Albishri, A. A., & Dessouky, M. M. (2024). Comparative analysis of machine learning techniques for phishing detection. *Engineering Technology and Applied Science Research*.
3. Sakhare, N. N., et al. (2024). Phishing website detection using machine learning techniques. *International Journal of Intelligent Systems*.
4. Ajjam, W. S., & Ibrahim, A. A. (2025). Phishing URL detection using LSTM networks. *Journal of Computer Science Research*.
5. Kibriya, H., et al. (2025). Lightweight malicious URL detection using deep learning. *Scientific Reports*.
6. Almujahid, N. F., et al. (2024). Machine learning based phishing site detection. *PeerJ Computer Science*.
7. Patil, M., et al. (2024). Deep neural network for malicious URL detection. *African Journal of Biomedical Research*.
8. Mundodagi, S., et al. (2024). Detection of phishing URLs using machine learning. *Journal of Web Engineering*.
9. Dubey, R., et al. (2025). Phishing detection using CNN models. *arXiv preprint*.
10. Hossain, M. I., et al. (2025). Graph-based LSTM model for malicious URL detection. *arXiv preprint*.