PHISHING WEBSITE DETECTION USING URL ANALYSIS

Ms. Sowmiya. S, M.Sc.,

Assistant Professor,
Department of Computer Science with Cyber Security,
Sri Ramakrishna College of Arts & Science,
Coimbatore-06.
Ph: 9025140467
ssowmiya@srcas.ac.in

Mr. Rajesh Kumar. B
III-BSc CS with Cyber Security
Department of Computer Science with Cyber Security,
Sri Ramakrishna College of Arts & Science, Coimbatore-06.
23130033@srcas.ac.in

Mr. Madesh. A
III-BSc CS with Cyber Security
Department of Computer Science with Cyber Security,
Sri Ramakrishna College of Arts & Science, Coimbatore-06.
23130022@srcas.ac.in

ABSTRACT

Phishing websites are one of the most common cyber threats in today's digital environment. These websites are designed to imitate legitimate websites in order to deceive users into providing sensitive information such as usernames, passwords, banking details, and other personal data. With the rapid increase in online services and digital transactions, identifying phishing websites at an early stage has become a major challenge in the field of cybersecurity.

This project focuses on detecting phishing websites using URL analysis, which is considered an efficient and lightweight method compared to traditional content-based detection techniques. Instead of analyzing the full webpage content, the system studies the structure and characteristics of the website's URL. Several features are extracted from the URL, including its length, the presence of special characters, the use of IP addresses instead of domain names, the number of subdomains, suspicious keywords, HTTPS usage, and redirection behavior. These characteristics help determine whether a website is legitimate or potentially malicious.

Machine learning techniques are used to classify URLs as either phishing or legitimate. A dataset consisting of both phishing and genuine URLs is collected and used for training and testing the model. After performing data preprocessing and feature extraction, a classification algorithm is trained to recognize patterns that are commonly associated with phishing URLs. Once trained, the model can analyze new or unknown URLs and predict whether they are safe or malicious with high accuracy.

The proposed system helps protect users from phishing attacks by offering quick and automated detection without the need to access the website's actual content. This approach makes the system safer and more efficient. Overall, the project demonstrates that combining URL-based analysis with machine learning provides a reliable method for detecting phishing websites and can be integrated into web browsers, email filtering systems, or other cybersecurity tools to improve online safety.

Keywords: Phishing Detection, URL Analysis, Machine Learning, Cybersecurity, Feature Extraction, Web Security.

1.INTRODUCTION

Phishing is a cyberattack technique where attackers create fake websites that closely resemble legitimate ones to steal sensitive information such as usernames, passwords, and banking details. These deceptive websites make users believe they are visiting trusted platforms, which allows attackers to collect confidential data. Because of this, phishing has become one of the most common and serious threats on the internet.

With the rapid growth of online services such as online banking, e-commerce, and social media platforms, phishing attacks have increased significantly in recent years. Traditional phishing detection methods mainly depend on blacklist databases to identify malicious websites. However, these approaches are often not fully effective because new phishing websites are created continuously and may not be immediately added to the blacklist.

To address this issue, modern phishing detection systems use URL analysis combined with machine learning techniques. In this method, various features of a URL—such as its length, the presence of special characters, and unusual or suspicious patterns—are examined to

determine whether the website is legitimate or potentially phishing. The goal of this project is to design a system that can efficiently detect phishing websites by analyzing URL characteristics, which helps strengthen online security and protect users from cyber fraud.

Furthermore, the proposed system helps users recognize suspicious websites before they actually visit them. By automatically analyzing URLs and classifying them as either safe or malicious, the system provides an additional layer of protection for internet users. This method not only improves the accuracy of phishing detection but also contributes to the development of more advanced cybersecurity solutions for safer online activities.

2. LITERATURE REVIEW

1.Author: Ma, J.

Title: **Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs**

Ma, J. introduced a method for identifying malicious and phishing websites by analyzing URL-based features instead of depending only on traditional blacklist systems. The study explains that blacklist approaches are limited because they cannot easily detect newly created phishing websites that have not yet been added to the database.

In this research, several characteristics of URLs—such as URL length, tokens within the URL, and suspicious patterns—are extracted and used as features. These features are then applied to train machine learning models that can classify websites as either legitimate or malicious. The proposed method is capable of detecting previously unknown phishing websites with a good level of accuracy.

The study also highlights that URL-based detection methods are suitable for real-time applications. By applying machine learning techniques, the system reduces the need for frequent manual updates to blacklist databases. Experimental results show that this approach performs better than many traditional detection methods.

Overall, this research demonstrates that analyzing URL characteristics is an effective strategy for detecting phishing websites, and it strongly supports the concept used in our project.

2. **Author: Gareca, S.**

Title: A Framework for Detection and Measurement of Phishing Attacks

Gareca, S. proposed a framework for detecting and analyzing phishing attacks by examining their common characteristics. The author explains that phishing websites are usually designed to closely resemble legitimate websites in order to trick users into revealing sensitive information.

The study mainly focuses on identifying key indicators of phishing, particularly those related to the structure of URLs and domain properties. Several important features are discussed in detail, such as misleading domain names, the presence of multiple or excessive subdomains, and unusual or suspicious URL patterns. These characteristics help in identifying whether a website may be involved in phishing activities.

The framework also helps in understanding the scale and impact of phishing attacks. In addition, the paper points out the limitations of traditional blacklist-based detection systems, which are often unable to identify newly created phishing websites.

By analyzing common phishing patterns, the study provides valuable insights for developing more effective phishing detection systems. The research highlights the importance of URL-based features in identifying phishing websites and helps in understanding the behavior and strategies used by attackers. Overall, this work supports the use of URL analysis as an effective approach for phishing website detection systems.

3. **Author: Abu-Nimeh S.**

Title: A Comparison of Machine Learning Techniques for Phishing Detection

Abu-Nimeh, S. carried out a comparative study to evaluate different machine learning algorithms used for detecting phishing websites. In this research, several models such as Naive Bayes, Decision Tree, and Support Vector Machine were tested using datasets that contained both phishing and legitimate website information.

The study explains that machine learning techniques can automatically learn patterns and characteristics that help distinguish phishing websites from genuine ones. To evaluate the effectiveness of the models, performance measures such as accuracy and error rate were used.

The results indicate that machine learning methods can achieve strong detection performance when appropriate features are selected. The research also discusses the advantages and limitations of each algorithm in the context of phishing detection.

Overall, the study emphasizes the importance of selecting suitable algorithms and relevant features to improve phishing detection systems. It also shows that machine learning can reduce manual effort and increase automation in identifying phishing websites. Compared to traditional detection methods, these approaches demonstrate better performance, which supports the use of machine learning in phishing website detection projects.

3. PROPOSED SYSTEM

The proposed system detects phishing websites by using URL analysis along with machine learning techniques. Instead of accessing and analyzing the full webpage content, the system focuses on extracting important features directly from the URL. These features include the length of the URL, the number of subdomains, the presence of special characters, and whether HTTPS is used.

After extracting these features, a trained machine learning model analyzes the URL and classifies it as either phishing or legitimate. This method allows the system to identify suspicious websites quickly without needing to load the actual webpage.

The approach provides a fast, secure, and accurate way to detect phishing websites. It can also be applied in real-time environments such as web browsers, email filtering systems, and other cybersecurity tools to improve online safety.

3.1 URL -Based Phishing Detection Approach

The system uses URL-based analysis to identify phishing websites without accessing or analyzing the webpage content. It extracts key features from the URL, such as its length, the number of subdomains, the presence of special characters, suspicious keywords, and the use of HTTPS. These features help capture patterns that are commonly found in phishing URLs.

This method is fast, lightweight, and secure, which makes it suitable for real-time phishing detection.

4. METHODOLOGY

4.1 System Implementation

The system implementation begins with collecting a dataset that contains both phishing and legitimate URLs. The collected data is then preprocessed to prepare it for analysis. After preprocessing, important features are extracted from each URL, such as the URL length, the presence of special characters, and other suspicious patterns.

These extracted features are used to train a machine learning model that can classify URLs as either phishing or legitimate. Once the training process is completed, the model is integrated into the system. The final system can then analyze new URLs and effectively detect whether they are phishing websites or safe ones.

- **Data Collection**

Data collection is the initial step in developing the phishing detection system. In this stage, both phishing and legitimate URLs are gathered from reliable public datasets and trusted online sources. The dataset contains labeled URLs, clearly indicating whether each website is phishing or legitimate.

Collecting a large and diverse dataset is important because it helps the model learn different patterns used in phishing attacks. The data includes URLs from various sectors such as banking, social media, and e-commerce. This variety helps improve the model's ability to generalize and detect phishing websites more effectively.

The dataset is carefully reviewed to ensure its quality and relevance. Any unreliable, duplicate, or outdated URLs are removed during this process. The collected dataset serves as the foundation for the entire machine learning process, and proper data collection plays a crucial role in improving the accuracy of the phishing detection system.

- **Data Preprocessing and Cleaning**

Data preprocessing is carried out to clean and prepare the collected URLs for further analysis. During this stage, duplicate URLs are removed to prevent bias during model training. Any invalid or corrupted URL entries are also identified and eliminated from the dataset.

The URLs are then converted into a standardized format to ensure consistent processing. Missing or incomplete values are properly handled to maintain the quality of the dataset. This preprocessing step helps reduce noise and improves the overall quality of the data.

In addition, the dataset is examined to check for any class imbalance between phishing and legitimate URLs. If necessary, simple balancing techniques are applied to maintain a fair distribution of data. Clean and well-prepared data allows the model to learn meaningful patterns more effectively. Proper preprocessing therefore improves the overall performance and reliability of the phishing detection system.

- **URL Feature**

URL feature extraction is the process of identifying and calculating important characteristics from each URL. In this step, several features are extracted, such as the length of the URL, the number of dots, the number of subdomains, and the presence of special characters. The system also checks whether an IP address is used instead of a domain name. In addition, the use of HTTPS is examined as a security indicator, and suspicious keywords within the URL are identified. These features help capture patterns that are commonly associated with phishing websites. Feature extraction transforms raw URLs into numerical feature vectors that can be used by machine learning models for classification.

- **Feature Selection and Encoding**

Feature selection involves identifying and choosing the most important features for training the model. During this process, irrelevant or redundant features are removed to reduce noise in the dataset. Selecting only the significant features helps improve the accuracy of the model and also reduces the computational time required for training.

Encoding is applied to convert categorical features into numerical values, since machine learning algorithms require numerical data to process information. In some cases, feature scaling or normalization is also performed to bring feature values into a similar range. This helps improve the stability and convergence of the model during training.

Proper feature selection helps reduce the risk of overfitting, while encoding ensures that the data is compatible with different machine learning algorithms. Overall, this step improves both the efficiency and performance of the phishing detection system.

- **Model Selection**

Model selection involves choosing the most suitable machine learning algorithm for detecting phishing websites. Different classification models are evaluated based on factors such as accuracy, efficiency, and reliability.

The chosen model should be capable of effectively handling URL-based features. In many cases, simpler models are preferred because they provide faster training and prediction times. The selection of the model plays an important role in determining the overall detection performance of the system.

The final model is chosen after analyzing preliminary experimental results. Factors such as training time, prediction speed, and overall accuracy are carefully considered. In addition, the model should be able to generalize well when analyzing new and unseen URLs. Ease of integration with the overall system is also taken into account.

Proper model selection helps maintain a good balance between system performance and computational complexity.

- **Model Training and Validation**

Model training involves providing the extracted features along with their corresponding labels to the selected machine learning algorithm. The dataset is first divided into two parts: a training set and a testing set. The model learns patterns and relationships from the training data.

After the training phase, the model is evaluated using the testing data to measure its performance. Different evaluation metrics such as accuracy and recall are calculated to assess how well the model can identify phishing and legitimate URLs.

To further improve performance, parameter tuning is carried out by adjusting the model's settings. During this process, overfitting is carefully monitored and controlled to ensure that the model performs well not only on training data but also on new and unseen URLs.

- **System Integration**

System integration involves connecting the trained machine learning model with the application's backend and user interface. When a user enters a URL, it is sent to the backend for processing. The backend then performs feature extraction on the given URL. After extracting the necessary features, the trained model analyzes the data and predicts whether the URL is phishing or legitimate.

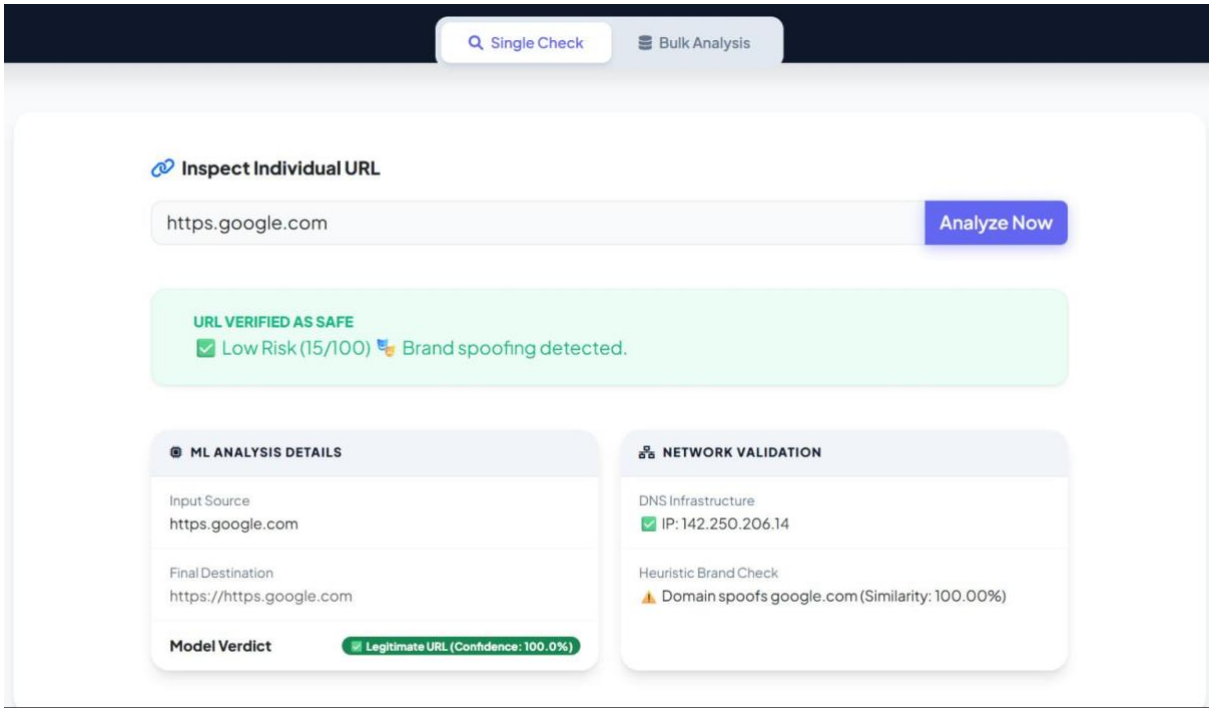
5. Results and Discussion

The proposed phishing website detection system was evaluated using a dataset that included both phishing and legitimate URLs. Different features related to the structure of URLs were extracted and used to train a machine learning model for classification. After training, the model was tested with separate data to assess its ability to correctly identify phishing websites.

The results indicate that the system can successfully classify URLs as either phishing or legitimate with a good level of accuracy. The model was able to identify most phishing URLs by analyzing suspicious patterns such as unusually long URLs, the presence of uncommon characters, and the use of IP addresses instead of domain names. Evaluation metrics including accuracy, precision, and recall were used to measure the performance of the model, and the findings show that URL-based analysis is an effective method for detecting phishing websites.

The discussion also highlights that the system offers a fast and efficient way to detect phishing websites without needing to analyze the complete webpage content. This makes the detection process quicker and suitable for real-time applications. However, the system may require

regular updates and larger datasets to further improve its accuracy as new phishing techniques continue to appear. Overall, the proposed approach contributes to improving cybersecurity and helps protect users from online fraud.



6. CONCLUSION

The phishing website detection system uses URL analysis and machine learning techniques to identify malicious websites. The system analyzes different URL features and classifies websites as either phishing or legitimate without accessing the actual webpage content. This approach ensures fast and secure detection.

The proposed method reduces the reliance on traditional blacklist-based detection systems and minimizes the need for manual verification. Experimental results show that the system can accurately detect phishing URLs in real-time environments.

The system is designed with a modular structure, which makes it easy to maintain and expand in the future. In addition, the simple user interface allows even non-technical users to easily check suspicious links. The project demonstrates the practical use of machine learning techniques in the field of cybersecurity.

It also emphasizes the importance of detecting phishing attacks at an early stage to prevent data theft and online fraud. The system helps improve user awareness by providing clear warnings when unsafe websites are detected. Overall, the project successfully meets its objectives and offers an effective solution for phishing website detection.

The implementation is lightweight and suitable for academic purposes as well as small-scale applications. The system design ensures both reliability and efficiency, and the results confirm the effectiveness of URL-based phishing detection. This work contributes to enhancing online security and can serve as a foundation for future research and development in cybersecurity applications.

7. REFERENCES

1." Phishing and Communication Channels: A Guide to Identifying and Mitigating Phishing Attacks (2022)","Gunikhan Sonowal","Apress (Springer Nature)"

https://link.springer.com/book/10.1007/978-1-4842-7744-7/Marchette/p/book/9781032401003

2 "Cybersecurity Analytics (2023)","Rakesh M. Verma"," Routledge"

https://www.routledge.com/Cybersecurity-Analytics/Verma

3. "Phishing: Detection, Analysis, and Prevention (2022)","Amrita Mitra","

(E-book/Print edition) "

https://www.amazon.in/Phishing-Detection-Prevention-Amrita-Mitra-ebook/dp/B0888PSB92

4." Tiny Machine Learning: Design Principles and Applications (2026)","Houbing Song","Wiley-IEEE Press"

Wiley-IEEE Info: (publisher link) search Wiley site.

5. “Applied Graph Data Science: Graph Algorithms and Platforms (2024)”,” Houbing Song”,”Elsevier”

6. “smart Transportation: AI Enabled Mobility and Autonomous Driving (relevant for AI security examples) (2025)”,”Houbing Song”,”CRC Press”

7. “Phishing Dark Waters: The Offensive and Defensive Sides of Malicious Emails (2nd Ed.)”,”Christopher Hadnagy”,” Wiley — anticipated updated editions by 2025 “Protecting and Mitigating Against Cyber Threats (O’Reilly)(2024)”,”O’Reilly Media”

8. “Neuro-symbolic AI: Foundations and Applications (2026)”,” Houbing Song”,”Wiley-IEEE Press ”

9. “Federated Learning for Digital Healthcare Systems (2024)”,”Houbing Song”,” Elsevier”