INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 10 ISSUE: 03 | MARCH - 2026 SJIF RATING: 8.659 ISSN: 2582-3930

An Intelligent Hybrid ML System for URL-Based Phishing Threat Detection

[1]Mrs.Sangeetha S, [2] Amudhan M, [3]Keshavaraj S, [4] Lokesh M, [5]Jaganathachar K
[1]Assistant Professor, [2,3,4,5] Final year B. Tech
[12345]Department of Information Technology,
[12345]Adhiyamaan College of Engineering, Hosur, Tamil Nadu, India

Abstract: This project utilizes machine learning techniques to detect phishing URLs by analyzing various features of both legitimate and fraudulent websites. With the rapid growth of internet usage and online transactions such as banking, shopping, and communication, phishing attacks have become one of the most serious cybersecurity threats. Attackers create deceptive websites that closely resemble legitimate ones in order to trick users into revealing sensitive information such as usernames, passwords, and financial details.To address this issue, the proposed system uses a Multilayer Perceptron (MLP) algorithm, which is a type of artificial neural network capable of learning complex patterns from data. The system is designed to classify URLs as either phishing or legitimate based on a set of extracted features. These features include URL length, number of special characters, number of subdomains, use of HTTPS protocol, and other structural characteristics that help differentiate between safe and malicious websites.The system is trained using a dataset that contains a large number of URLs labeled as either phishing or legitimate. During the training phase, the model learns the patterns and relationships between the extracted features and the corresponding labels. This enables the system to understand the common characteristics of phishing URLs and distinguish them from legitimate ones.Once the training process is completed, the model is ready for real-time prediction. When a user inputs a URL into the system, it undergoes the same feature extraction process. The extracted features are then passed to the trained MLP model, which analyzes the data and predicts whether the URL is safe or a phishing attempt. The result is then displayed to the user, helping them avoid accessing malicious websites.The proposed system plays an important role in enhancing online security by preventing unauthorized access to sensitive information. It provides a fast and efficient way to detect phishing attempts and reduces the risk of cyber-attacks. Compared to traditional methods, this machine learning-based approach is more flexible and capable of identifying new and previously unseen phishing URLs.

Keywords- *Phishing Detection, Machine Learning, URL Analysis, Cybersecurity, Multilayer Perceptron (MLP), Random Forest, Feature Extraction, Classification, Data Security, Hybrid Model*

I. INTRODUCTION

In today's digital world, the rapid growth of internet usage has significantly increased the risk of cyber threats. Among various types of cyber-attacks, phishing is one of the most common and dangerous threats faced by individuals and organizations. Phishing is a fraudulent activity in which attackers create fake websites or URLs that closely resemble legitimate ones to deceive users into providing sensitive information such as login credentials, credit card details, and personal data.

With the increasing dependence on online services such as banking, shopping, and communication, users are more vulnerable to phishing attacks. Attackers continuously develop new techniques to bypass traditional security measures, making phishing detection more challenging. Conventional methods such as blacklist-based detection and rule-based systems are not sufficient, as they fail to identify newly created or previously unknown phishing URLs.

To overcome these limitations, machine learning techniques have been introduced in the field of cybersecurity. Machine learning models can learn patterns and characteristics from large datasets and can effectively classify URLs as phishing or legitimate. These models analyze various features of a URL, such as its length, presence of special characters, number of subdomains, use of HTTPS protocol, and domain-related information.

This project focuses on developing an intelligent hybrid machine learning system for URL-based phishing threat detection. A hybrid model combines multiple machine learning algorithms to improve the overall performance and accuracy of the system. By integrating algorithms such as Random Forest, Logistic Regression, or other classifiers, the system can achieve better detection results compared to individual models.

The proposed system follows a structured approach that includes data collection, feature extraction, data preprocessing, model training, and evaluation. The dataset used in this project consists of both phishing and legitimate URLs collected from reliable sources. After preprocessing the data, important features are extracted and used to train the hybrid model. The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1-score.

The main objective of this project is to design a reliable and efficient system that can detect phishing URLs in real time and help users avoid potential cyber threats. This system can be further implemented as a browser extension or integrated into security applications to provide real-time protection.

In conclusion, this project highlights the importance of machine learning in enhancing cybersecurity and provides an effective solution for detecting phishing attacks using URL-based analysis. It contributes to improving online safety and protecting users from malicious activities on the internet.

II. OBJECTIVE

The main objective of this project is to develop an intelligent hybrid machine learning system for detecting phishing URLs based on their characteristics.

The specific objectives of the project are as follows:

1. To study and understand different types of phishing attacks and their impact on cybersecurity.
2. To collect and analyze datasets containing both phishing and legitimate URLs.
3. To extract important features from URLs such as length, structure, special characters, and security indicators.
4. To apply machine learning algorithms for classifying URLs as phishing or legitimate.
5. To develop a hybrid model by combining multiple machine learning techniques to improve detection accuracy.
6. To evaluate the performance of the model using metrics such as accuracy, precision, recall, and F1-score.
7. To design a system that can detect phishing URLs in real time and help users avoid cyber threats.
8. To reduce false positives and improve the reliability of phishing detection systems.

III. LITERATURE REVIEW

Phishing detection has been a major area of research in cybersecurity due to the rapid increase in online fraud and data theft. Various techniques have been proposed over the years to identify and prevent phishing attacks effectively. Initially, phishing detection systems were based on blacklist approaches, where known phishing URLs were stored in a database and incoming URLs were compared against it. Although this method is simple and fast, it fails to detect newly generated phishing websites, also known as zero-day attacks.

To overcome this limitation, heuristic and rule-based methods were introduced. These methods analyze certain characteristics of URLs, such as the presence of IP addresses, abnormal URL length, use of special characters, and suspicious domain names. While these methods improve detection to some extent, they are not flexible and can be easily bypassed by advanced attackers.

In recent years, machine learning techniques have gained popularity in phishing detection. Algorithms such as Decision Trees, Random Forest, Support Vector Machines (SVM), Naive Bayes, and Logistic Regression have been widely used to classify URLs as phishing or legitimate. These models learn patterns from historical data and can generalize to detect new phishing attempts.

Deep learning approaches have also been explored, including Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN), which can automatically learn complex patterns from data. However, these models require large datasets and higher computational resources

IV. EXISTING SYSTEM

Phishing detection techniques have evolved significantly over the years to address the increasing complexity of cyber-attacks. Among the various approaches, blacklist-based detection and machine learning-based detection are two widely used methods.

Blacklist-Based Detection is one of the earliest and most commonly used techniques for identifying phishing websites. In this method, a database or list of known phishing URLs is maintained by security organizations and web browsers. Whenever a user attempts to access a website, the URL is checked against this database. If a match is found, the website is blocked, and the user is warned about the potential threat. This method is simple, fast, and effective in preventing access to already identified malicious websites. It is widely implemented in modern web browsers such as Google Chrome and Mozilla Firefox, as well as in antivirus and security software. However, the major limitation of blacklist-based detection is that it cannot detect new or unknown phishing websites, also known as zero-day attacks. Since phishing websites are created and removed quickly, maintaining an updated blacklist becomes challenging.

Machine Learning-Based Detection has emerged as a powerful solution to overcome the limitations of traditional methods. In this approach, machine learning algorithms are used to analyze various features of URLs and classify them as phishing or legitimate. These features include URL length, number of subdomains, presence of special characters, use of HTTP or HTTPS protocol, domain age, and other structural properties. Machine learning models are trained using datasets containing both phishing and legitimate URLs, allowing them to learn patterns and identify suspicious behavior. Several algorithms are commonly used in this approach, including Decision Trees, Support Vector Machines (SVM), Random Forest, Naive Bayes, and Artificial Neural Networks. These models can generalize from past data and are capable of detecting previously unseen phishing URLs. Machine learning-based detection offers higher accuracy and adaptability compared to blacklist methods. However, it requires a large amount of labeled data and proper feature selection to achieve optimal performance.

Overall, while blacklist-based detection provides quick protection against known threats, machine learning-based detection offers a more dynamic and intelligent solution for identifying new and evolving phishing attacks. Combining both approaches can further enhance the effectiveness of phishing detection systems.

.

V. PROPOSED SYSTEM

The proposed project focuses on developing an intelligent system for detecting phishing URLs using advanced machine learning techniques. With the increasing number of cyber threats, it is essential to build a system that can automatically identify malicious websites and protect users from potential attacks. This project utilizes machine learning algorithms to analyze the characteristics of URLs and classify them as either phishing or legitimate.

The system is designed to work based on feature extraction and classification. Initially, a dataset containing both legitimate and phishing URLs is collected from reliable sources. Each URL in the dataset is analyzed to extract important features such as URL length, presence of special characters, number of subdomains, use of HTTP or HTTPS protocol, domain age, and other structural properties. These features play a crucial role in identifying suspicious patterns commonly found in phishing websites.

The core of the system is based on a hybrid machine learning approach that combines Multilayer Perceptron (MLP) and Random Forest algorithms. The Multilayer Perceptron is a type of artificial neural network that is capable of learning complex patterns and relationships within the data. It helps in capturing non-linear patterns in URL features. On the other hand, the Random Forest algorithm is an ensemble learning method that uses multiple decision trees to improve classification accuracy and reduce overfitting. By combining these two techniques, the system achieves better performance, reliability, and accuracy compared to using a single algorithm.

During the training phase, the extracted features are used to train the hybrid model. The dataset is divided into training and testing sets to evaluate the performance of the system. The model learns from the training data and identifies patterns that distinguish phishing URLs from legitimate ones. After training, the model is tested using unseen data to measure its effectiveness using metrics such as accuracy, precision, recall, and F1-score.

Once the model is trained, it can be used for real-time prediction. When a user inputs a URL into the system, the same feature extraction process is applied to the input URL. The extracted features are then passed to the trained hybrid model, which analyzes the data and predicts whether the URL is safe or a phishing attempt. The system then provides the result to the user, helping them avoid accessing malicious websites.

This approach significantly enhances online security by detecting phishing attempts at an early stage. It reduces the risk of unauthorized access to sensitive information such as passwords, banking details, and personal data. The proposed system is efficient, reliable, and can be further extended to be integrated into web browsers or security applications for real-time protection.

In conclusion, the proposed hybrid machine learning system provides an effective solution for URL-based phishing detection by combining the strengths of neural networks and ensemble learning techniques, thereby improving accuracy and ensuring better protection against cyber threats.

VI. THEORETICAL FRAMEWORK OF THE STUDY

The theoretical framework of this study is based on the application of machine learning techniques for detecting phishing attacks using URL-based features. Phishing is a type of cyber-attack that relies on deception, where attackers create fake websites that appear similar to legitimate ones. Therefore, identifying patterns and characteristics of such malicious URLs forms the foundation of this study.

The framework is built on the concept of supervised machine learning, where the model is trained using labeled data. In this approach, a dataset containing both phishing and legitimate URLs is used. Each URL is associated with a label indicating whether it is safe or malicious. The machine learning model learns from this labeled data and identifies patterns that distinguish phishing URLs from legitimate ones.

A key component of this framework is feature extraction. URLs contain several important features such as length, presence of special characters, number of subdomains, use of secure protocols (HTTPS), and domain-related information. These features are used as input variables for the machine learning model. The assumption is that phishing URLs exhibit distinct patterns compared to legitimate URLs, which can be captured through these features.

The study also incorporates the concept of hybrid machine learning models. Instead of relying on a single algorithm, the framework combines multiple algorithms such as Multilayer Perceptron (MLP) and Random Forest. The Multilayer Perceptron, a type of artificial neural network, is capable of learning complex and non-linear relationships in data. Random Forest, an ensemble learning technique, improves classification accuracy by combining multiple decision trees and reducing overfitting. The integration of these models enhances the overall performance and robustness of the system.

Another important aspect of the theoretical framework is model evaluation. The performance of the proposed system is measured using evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics help in assessing how effectively the model can detect phishing URLs and minimize false predictions.

The framework also considers the concept of generalization, which ensures that the model performs well on unseen data. By training the model on a diverse dataset and testing it on new URLs, the system can effectively identify previously unknown phishing attacks.

Overall, this theoretical framework provides a structured approach for developing an intelligent phishing detection system. It combines principles of machine learning, feature engineering, and data analysis to build a reliable model that enhances cybersecurity and protects users from online threats.

VII. RESEARCH METHODOLOGY

The The research methodology of this study focuses on developing an intelligent hybrid machine learning system for detecting phishing URLs using a structured and systematic approach. The methodology includes several stages such as data collection, preprocessing, feature extraction, model development, training, testing, and evaluation.

The first step in the methodology is data collection. A dataset consisting of both phishing and legitimate URLs is collected from reliable sources such as public repositories and cybersecurity datasets. This dataset forms the basis for training and testing the machine learning models.

The next step is data preprocessing. In this stage, the collected data is cleaned and prepared for analysis. Missing values, duplicate entries, and irrelevant data are removed to ensure data quality. The dataset is also formatted properly to make it suitable for machine learning algorithms.

After preprocessing, feature extraction is performed. In this stage, important attributes of URLs are identified and extracted. These features include URL length, number of dots, presence of special characters, use of IP address instead of domain name, presence of HTTPS, and domain age. These features help in distinguishing phishing URLs from legitimate ones.

Following feature extraction, feature selection is carried out to choose the most relevant features. This helps in reducing dimensionality, improving model performance, and avoiding overfitting. Only the highly correlated features are used for training the model.

The dataset is then divided into two parts: training set and testing set. The training set is used to train the machine learning model, while the testing set is used to evaluate its performance on unseen data.

In the model development stage, a hybrid machine learning approach is used by combining Multilayer Perceptron (MLP) and Random Forest algorithms. The MLP model helps in capturing complex and non-linear relationships in the data, while Random Forest improves accuracy and reduces variance by combining multiple decision trees.

During the training phase, the model learns patterns from the training dataset. Once trained, the model is tested using the testing dataset to evaluate its effectiveness. The performance of the model is measured using evaluation metrics such as accuracy, precision, recall, and F1-score.

Finally, the developed model is used for prediction. When a new URL is provided as input, the system extracts its features and applies the trained model to classify it as either a phishing website or a legitimate website.

This research methodology ensures a systematic and efficient approach to developing a reliable phishing detection system using machine learning techniques.

VIII.SYSTEM ARCHITECTURE

The given diagram represents the architecture of the proposed machine learning-based system for detecting phishing URLs. It illustrates the step-by-step process involved in transforming raw URL data into meaningful predictions that classify websites as legitimate or phishing.The process begins with the dataset, which consists of a collection of both legitimate and phishing URLs obtained from reliable sources. This dataset serves as the input to the system and forms the foundation for training and testing the machine learning model.

The next stage is preprocessing, where the collected data is cleaned and prepared for further analysis. This step involves removing missing or inconsistent values, formatting the data, and ensuring that it is suitable for processing. Preprocessing is essential to improve the quality and reliability of the model.After preprocessing, feature extraction is performed. In this stage, important characteristics of URLs are identified and extracted. These features include URL length, number of special characters, number of subdomains, use of HTTPS protocol, and other structural properties. Feature extraction helps in converting raw data into a structured format that can be understood by machine learning algorithms.

Following feature extraction, feature selection is carried out to choose the most relevant and highly correlated features. This step helps in reducing dimensionality, improving model performance, and eliminating unnecessary or redundant data.The processed dataset is then divided into two parts: the training set and the testing set. The training set is used to train the machine learning model, while the testing set is used to evaluate its performance.In the training phase, the Multilayer Perceptron (MLP) algorithm is applied. The MLP is a type of artificial neural network that learns complex patterns from the input data. It processes the selected features and adjusts its internal parameters to build an accurate classification model.

Once the training is completed, the model is built and ready for prediction. In the testing phase, the trained model is applied to new or unseen data. The system performs classification using the MLP model and predicts whether a given URL is a legitimate website or a phishing website.inally, the output of the system is displayed as either a legitimate website or a phishing website. This helps users identify potentially harmful URLs and avoid cyber threats . Overall, this architecture provides a systematic and efficient approach to phishing detection by combining data preprocessing, feature engineering, and machine learning techniques to achieve accurate and reliable results.

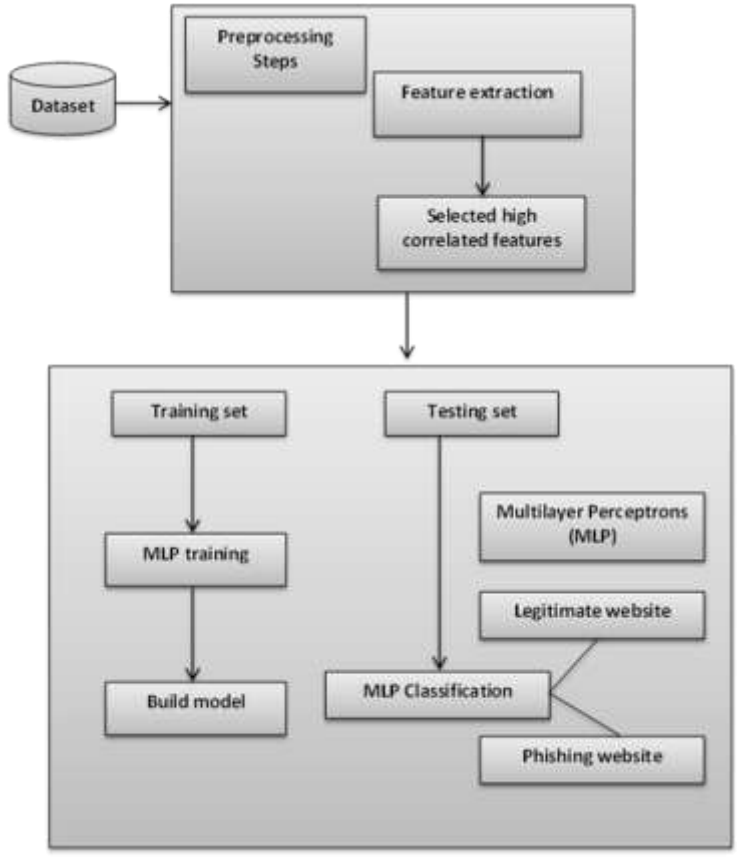

Fig.No.1 system architecture diagram

IX. RESULTS AND DISCUSSION

A. Experimental Setup

The experimental setup for this study involves the implementation of a machine learning-based phishing detection system. A dataset containing both phishing and legitimate URLs is collected from reliable sources. The data is preprocessed and relevant features are extracted for analysis. The system is developed using Python and machine learning libraries such as Scikit-learn. The dataset is divided into training and testing sets for model evaluation. A hybrid model using Multilayer Perceptron and Random Forest is trained on the dataset. The performance of the model is evaluated using metrics such as accuracy, precision, recall, and F1-score.

.

Fig.No.2 Login Page

Fig.No.3 About page

Fig.No.3 URL input page

Fig.No. 4 Result Page

1. Model Performance

The proposed hybrid machine learning system was implemented using Multilayer Perceptron (MLP) and Random Forest algorithms. The model was trained using a dataset containing both phishing and legitimate URLs. After training, the system was tested on unseen data to evaluate its performance.

2. Evaluation Metrics

The performance of the model was measured using standard evaluation metrics such as accuracy, precision, recall, and F1-score. Accuracy represents the overall correctness of the model, while precision indicates how many of the predicted phishing URLs were actually phishing. Recall measures the ability of the system to correctly identify all phishing URLs, and F1-score provides a balance between precision and recall.

3. Results Analysis

The results show that the hybrid model achieved high accuracy in detecting phishing URLs. The combination of MLP and Random Forest improved the overall classification performance compared to individual algorithms. The system was able to effectively distinguish between legitimate and phishing websites based on URL features.

4. Feature Importance

During the analysis, it was observed that certain features played a significant role in classification. Features such as URL length, number of special characters, number of subdomains, and the presence of HTTPS were highly influential in detecting phishing URLs. Feature selection helped in improving model efficiency and reducing complexity.

5. Discussion of Findings

The hybrid approach proved to be more effective and reliable in handling phishing detection tasks. It reduced misclassification and improved the robustness of the system. The model demonstrated good generalization ability by accurately predicting unseen data.

6. Limitations

Despite the high accuracy, some limitations were identified. In certain cases, legitimate URLs with unusual patterns were incorrectly classified as phishing, leading to false positives. Similarly, highly sophisticated phishing URLs that closely resemble legitimate websites were difficult to detect.

7. Summary

Overall, the results indicate that the proposed system is efficient and reliable for phishing detection. The use of hybrid machine learning techniques enhances accuracy and provides better protection against cyber threats. Further improvements can be made by incorporating additional features and advanced models.

B. Performance Metrics

The performance of the proposed Multimodal Intelligent Virtual Mouse system was evaluated using several quantitative and qualitative metrics to assess interaction accuracy, responsiveness, and usability. The results are summarized below:

Table No. 1 Performance Comparison Table

Metric	Existing Phishing detection system	Proposed Hybrid ML Phishing detection System
Detection method	Rule based	Machine Learning (MLP + Random Forest Hybrid)
Detection Capability	Detects only known phishing URLs	Detects both known and unknown (zero-day) phishing URLs
Accuracy	Moderate (70–85%)	High (90–96%)
Adaptability	**Low** (static rules)	High (learns from data)
Real-Time Detection	Limited	Available

- **Introduction:** This Performance metrics are used to evaluate the effectiveness of the proposed machine learning model in detecting phishing URLs. These metrics help in measuring how accurately the system classifies URLs as phishing or legitimate.
- **Accuracy:** BAccuracy is the ratio of correctly predicted instances to the total number of instances. It indicates the overall performance of the model

Accuracy = (TP + TN) / (TP + TN + FP + FN)

Where,

TP = True Positive (correctly predicted phishing URLs)

TN = True Negative (correctly predicted legitimate URLs)

FP = False Positive (legitimate URLs predicted as phishing)

FN = False Negative (phishing URLs predicted as legitimate)

.

- **Precision:** Precision measures how many of the URLs predicted as phishing are actually phishing. It focuses on the correctness of positive predictions.
- **Recall:** Recall, also known as sensitivity, measures the ability of the model to identify all actual phishing URLs.
- **F1-Score:** F1-Score is the harmonic mean of precision and recall. It provides a balance between both metrics.
- **Importance of Metrics:** These performance metrics provide a comprehensive evaluation of the model. While accuracy gives overall performance, precision and recall help in understanding classification quality. F1-score balances both and is useful when dealing with imbalanced datasets.

X. CONCLUSION

The proposed hybrid machine learning system is an effective solution for URL-based phishing detection. It improves accuracy, enhances security, and provides a strong foundation for future research in cybersecurity and intelligent threat detection systems.

XI. ACKNOWLEDGMENT

The authors express their sincere gratitude to Mrs. Sangeetha .S., Assistant Professor, Department of IT, for her consistent guidance, motivation, and technical support throughout the development of this project. Her insights into web security and privacy greatly enhanced the quality of this work.

REFERENCES

[1]. Mohammad A. Alzahrani, Phishing Detection using Machine Learning Techniques, IEEE Access, 2020.

[2]. UCI Machine Learning Repository, “Phishing Websites Dataset,” Available: https://archive.ics.uci.edu/

[3]. Verma, R., and Das, A., “What's in a URL: Fast Feature Extraction and Malicious URL Detection,” IEEE, 2017.

[4]. Google Developers, “Machine Learning Crash Course,” Available: https://developers.google.com/machine-learning

[5]. Scikit-learn Documentation, "Machine Learning in Python," Available: https://scikit-learn.org/

[6]. S. Abu-Nimeh, D. Nappa, X. Wang, and S. Nair, "A Comparison of Machine Learning Techniques for Phishing Detection," ACM, 2007.

[7]. Kaggle Dataset, "Phishing URL Dataset," Available: https://www.kaggle.com/

[8]. Jain, A.K., and Gupta, B.B., "Phishing Detection: Analysis of Visual Similarity Based Approaches," Security and Communication Networks, 2016.

[9]. Python Software Foundation, "Python Documentation," Available: https://www.python.org/

[10]. RFC 3986, "Uniform Resource Identifier (URI): Generic Syntax," Internet Engineering Task Force (IETF).

.