INTELLIGENT PHISHING URL DETECTION SYSTEM USING MACHINE LEARNING

N.V.Ashok Kumar[1], U.Vamsi[2], N.Aswini[3], K.Sai Sampath[4], M.Nirmala Devi[5]
Associate Professor[1], Student[2,3,4,5]
Department of Computer Science & Engineering(Data Science)[1,2,3,4,5]Avanthi institute of engineering and technology,Anakapalli,Andhra pradesh,India
,nagamashok@gmail.com[1],vamsiullingala@gmail.com[2], aswininimmadala@gmail.com[3]}
rkillada913@gmail.com[4], nirmalamalla133@gmail.com[5]} @aiet.ac.in

ABSTRACT

The rapid growth of internet usage has significantly increased cyber threats, particularly phishing attacks that aim to steal sensitive user information through malicious websites. Phishing URLs are often designed to closely imitate legitimate websites in order to deceive users and gain unauthorized access to credentials, financial data, and personal information. This paper presents an intelligent phishing URL detection and domain verification system using machine learning techniques. The proposed system incorporates domain validation mechanisms such as DNS resolution and WHOIS analysis to verify the authenticity of domains before classification. Following validation, relevant URL-based features are extracted and processed using multiple machine learning algorithms, including Logistic Regression, Random Forest, and Support Vector Machine (SVM). These models are trained and evaluated on real-world phishing datasets using performance metrics such as accuracy, confusion matrix, and classification report. The system is implemented using Python and deployed through a Streamlit-based web interface to enable real-time prediction. Experimental results demonstrate that the proposed system achieves high accuracy in detecting phishing URLs while providing confidence scores for predictions. This approach enhances cybersecurity awareness and helps prevent users from accessing potentially malicious websites.

INTRODUCTION

Digital transformation has significantly increased the use of online platforms, enabling users to perform transactions, communication, and data sharing over the internet with greater convenience and efficiency. However, this rapid growth has also led to a substantial rise in cyber threats, particularly phishing attacks. Phishing is a deceptive technique in which

attackers create fraudulent websites that closely resemble legitimate platforms to trick users into revealing sensitive information such as login credentials, banking details, and personal data. Traditional phishing detection methods primarily rely on blacklist-based approaches, which store known malicious URLs and block access to them. Although effective for previously identified threats, these methods are inadequate for detecting newly generated or unknown phishing URLs, making them less reliable in dynamic environments.

Existing systems attempt to identify malicious links using various techniques; however, they often lack real-time verification and fail to achieve high classification accuracy. To address these limitations, this paper proposes an intelligent phishing URL detection system integrated with domain verification techniques such as DNS resolution and WHOIS analysis. The proposed system extracts relevant URL features and applies multiple machine learning algorithms to accurately classify URLs as phishing or legitimate. Furthermore, the system is implemented using a Streamlit-based interface, enabling real-time predictions and user-friendly interaction. This approach enhances detection accuracy, improves reliability, and strengthens overall user security against phishing attacks.

LITERATURE REVIEW

This section reviews key prior works in phishing detection, analyzes existing techniques, and identifies the research gap that motivates this work. Jain and Gupta (2018) proposed a phishing detection approach using machine learning algorithms such as Logistic Regression and Decision Tree, focusing on extracting URL-based features to classify phishing websites and achieving improved accuracy compared to traditional blacklist methods. Abdelhamid et al. (2014) introduced an intelligent phishing detection system that analyzes both website content and URL features, demonstrating the effectiveness of classification algorithms in identifying phishing attacks. Similarly, Ma et al. (2009) developed a system for detecting malicious URLs using lexical and host-based features, emphasizing the importance of URL structure analysis and large-scale datasets for training robust models.

Zhang et al. (2011) proposed a real-time phishing detection system using Support Vector Machine (SVM), which achieved high detection rates but required continuous updates to handle evolving attack patterns. Verma and Das (2017) applied Random Forest and other ensemble learning techniques, showing that ensemble models outperform individual

classifiers in terms of accuracy and robustness. Furthermore, Sahingoz et al. (2019) proposed a machine learning-based phishing detection system that utilizes natural language processing and advanced feature extraction methods to analyze textual and structural patterns in URLs, thereby improving classification performance. Recent studies have also explored domain-based verification techniques such as DNS resolution and WHOIS analysis to validate domain authenticity before classification. These approaches help identify suspicious or inactive domains and reduce false positives, thereby enhancing detection reliability.

Despite these advancements, several research gaps remain. Most existing approaches rely either on machine learning models or blacklist-based methods without integrating domain verification mechanisms. As a result, they often fail to detect newly generated phishing URLs and lack real-time validation of domain authenticity. Additionally, many systems require large datasets and do not provide user-friendly interfaces for practical deployment. To address these limitations, the proposed work integrates DNS and WHOIS-based domain verification with multiple machine learning models, enabling accurate, efficient, and real-time phishing detection through a lightweight and user-friendly web interface.

METHODOLOGY

The proposed system follows a structured methodology to accurately detect phishing URLs using machine learning techniques combined with domain verification mechanisms. The dataset used in this work consists of both phishing and legitimate URLs collected from publicly available sources such as the UCI Machine Learning Repository and other open phishing datasets. The dataset includes a balanced distribution of malicious and benign URLs to ensure unbiased model training. For effective evaluation, the dataset is divided into three subsets: 70% for training, 15% for validation, and 15% for testing, allowing the models to generalize well on unseen data.

During the preprocessing stage, the collected URLs are cleaned and standardized by removing duplicate entries, handling missing values, and normalizing the data format. Feature extraction is then performed to generate meaningful attributes that can help distinguish phishing URLs from legitimate ones. These features include URL length, number of dots, presence of special characters such as '@' and '-', usage of HTTPS, and domain age. In addition to feature extraction, domain verification techniques such as DNS resolution and

WHOIS analysis are applied to validate whether a domain is active, registered, and trustworthy, thereby enhancing the reliability of the system.

The system employs multiple machine learning models, including Logistic Regression, Random Forest, and Support Vector Machine (SVM), to perform URL classification. Each model is trained using the extracted features, enabling it to learn patterns and relationships that differentiate phishing URLs from legitimate ones. The use of multiple models not only improves robustness but also allows comparative performance evaluation across different algorithms.

During the training phase, the models are optimized using appropriate hyperparameters to achieve better performance. The effectiveness of each model is evaluated using standard performance metrics such as accuracy, precision, recall, and F1-score. A confusion matrix is also used to analyze classification outcomes in detail. Finally, the entire system is implemented using Python and deployed through a Streamlit-based web interface, enabling real-time URL prediction and providing an interactive platform for users to analyze and verify URLs efficiently.

A. System Architecture

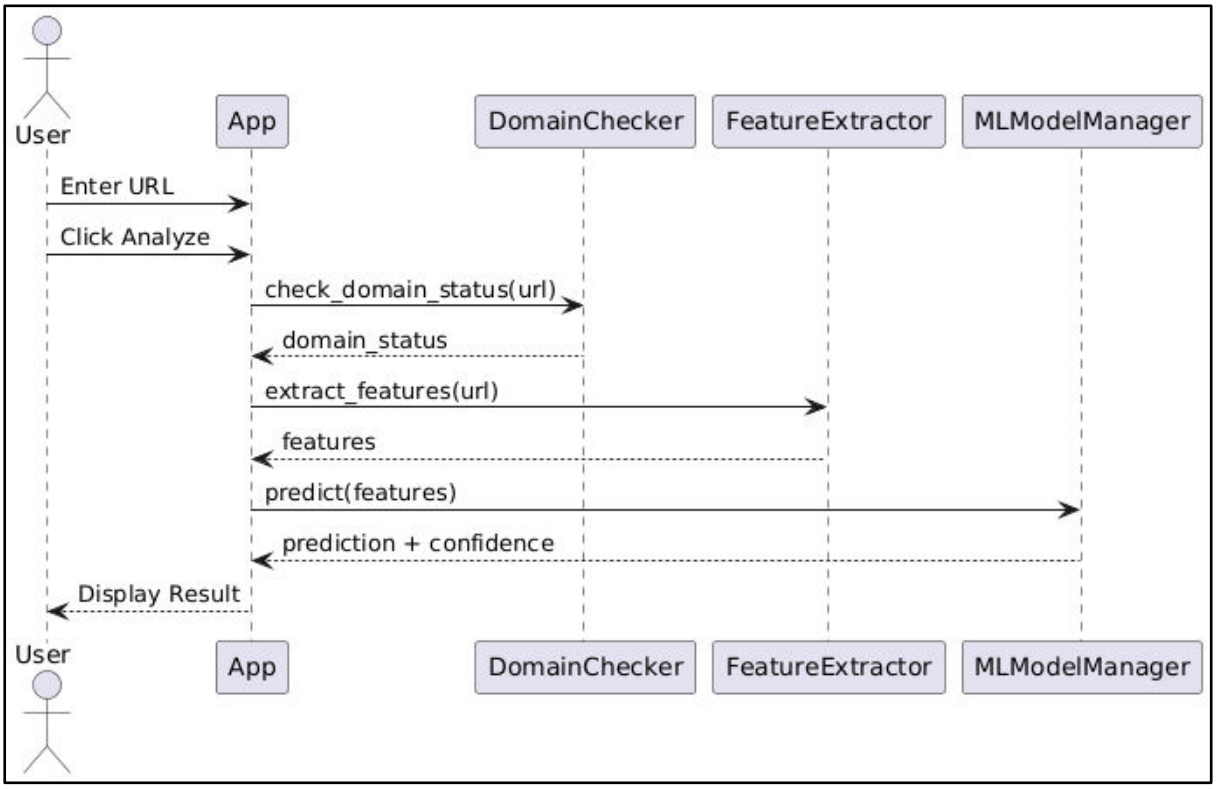
The proposed system follows a pipeline-based architecture consisting of six sequential stages to ensure accurate and efficient phishing URL detection. Initially, in the **URL Input stage**, the user provides a URL through a web-based interface for analysis. This input is then processed in the **Domain Verification stage**, where the system performs DNS resolution and WHOIS analysis to verify whether the domain is active, properly registered, and valid. This step helps in identifying suspicious or inactive domains before further processing.

In the next stage, **Feature Extraction**, the system extracts important URL-based features such as URL length, number of dots, presence of special characters (e.g., '@', '-'), HTTPS usage, and domain age. These features capture both structural and lexical characteristics of the URL, which are essential for effective classification. Following this, the **Data Preprocessing stage** normalizes and formats the extracted features to make them suitable for machine learning model input.

The processed data is then passed to the **Phishing Classification stage**, where multiple machine learning models, including Logistic Regression, Random Forest, and Support Vector

Machine (SVM), are applied to classify the URL as either phishing or legitimate. Each model independently analyzes the features and contributes to the final decision. Finally, in the **Output Visualization stage**, the system displays the prediction results along with confidence scores and model comparisons through a Streamlit-based web interface, enabling easy user interaction and understanding.

Overall, the system integrates domain verification techniques with machine learning-based classification to enhance detection reliability. The feature extraction module effectively captures key URL properties, while the classification module ensures accurate predictions using trained models. The Streamlit interface serves as the presentation layer, providing real-time input handling and clear visualization of results, thereby improving usability and decision-making.



Algorithm: Intelligent Phishing URL Detection using Machine Learning

Step 1: Accept the input URL UUU from the user through the web interface.

Step 2: Perform domain verification:

- Execute DNS resolution to check whether the domain exists.
- Perform WHOIS analysis to verify domain registration details.

Step 3: If the domain is invalid or inactive, classify the URL as *suspicious* and proceed to the output stage.

Step 4: Extract relevant URL-based features from UUU, including:

- URL length
- Number of dots
- Presence of special characters (e.g., '@', '-')
- HTTPS usage
- Number of subdomains

Step 5: Preprocess the extracted features by normalizing and converting them into a structured format suitable for model input.

Step 6: Load the trained machine learning models:

- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)

Step 7: For each model, compute the prediction PiP_iPi for the given URL.

Step 8: Aggregate the predictions from all models and calculate confidence scores for each classification.

Step 9: Determine the final classification
C∈{Phishing,Legitimate}C \in \{\text{Phishing}, \text{Legitimate}\}C∈{Phishing,Legitimate}
based on the combined outputs of the models (e.g., majority voting).

Step 10: Display the final prediction result, confidence scores, and model comparison through the Streamlit-based interface.

System Modules Description

The proposed intelligent phishing URL detection system is composed of several functional modules, each responsible for a specific task in the detection pipeline. The **Input Module** accepts the URL entered by the user through the Streamlit-based web interface. It ensures that the provided URL is in a valid format and prepares it for further processing, thereby reducing errors in subsequent stages.

The **Preprocessing Module** performs cleaning and normalization of the input URL. It removes unnecessary characters, handles inconsistencies, and standardizes the URL format to make it suitable for feature extraction. This step ensures that the data fed into the system is accurate and consistent.

The **Feature Extraction Module** plays a crucial role by extracting relevant lexical and structural features from the URL. These features include URL length, number of dots, presence of special characters such as '@' and '-', HTTPS usage, and the number of subdomains. These attributes help in distinguishing phishing URLs from legitimate ones and serve as inputs to the machine learning models.

The **Machine Learning Module** applies multiple trained algorithms, including Logistic Regression, Random Forest, and Support Vector Machine (SVM), to classify the URL. Each model independently processes the extracted features and generates predictions based on learned patterns.

The **Prediction Module** aggregates the outputs from all models and determines the final classification result. It also computes confidence scores for each prediction, which improves the reliability and interpretability of the system. By combining multiple model outputs, the system achieves better accuracy and robustness.

Finally, the **Visualization and Output Module** presents the classification result (phishing or legitimate) along with confidence scores and model comparisons through a Streamlit-based

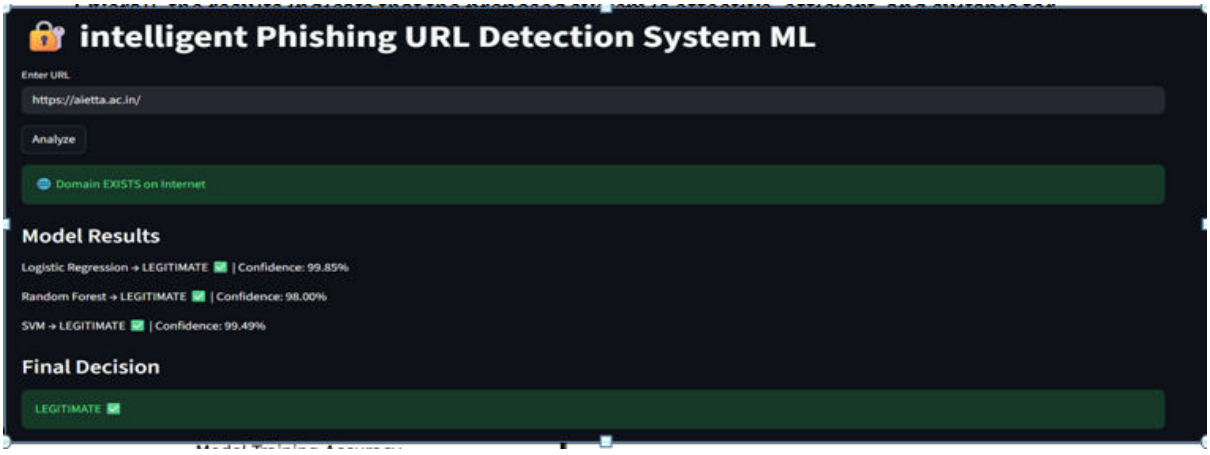
RESULTS AND DISCUSSION

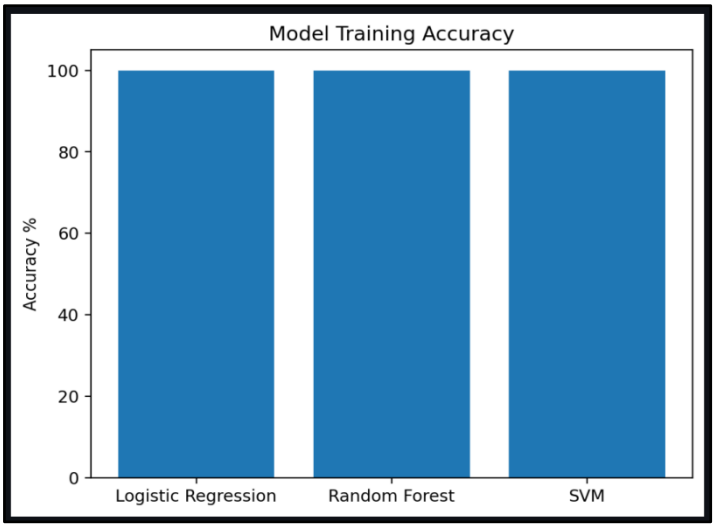
The experimental results demonstrate that the proposed system effectively identifies phishing URLs with high accuracy and reliability. The use of multiple machine learning models, including Logistic Regression, Random Forest, and Support Vector Machine (SVM), enables the system to achieve robust performance across different types of URLs. Among these models, Random Forest generally shows higher accuracy due to its ensemble learning capability, while SVM provides strong classification boundaries for complex patterns.

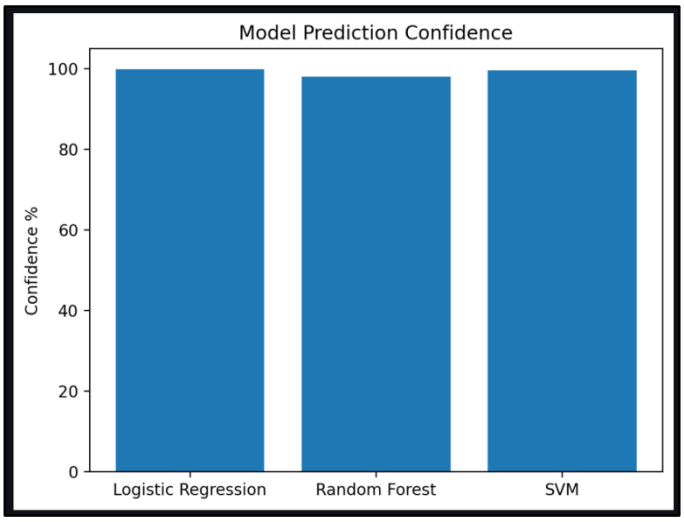
Performance evaluation is conducted using standard metrics such as accuracy, precision, recall, and F1-score. The confusion matrix analysis indicates that the system achieves a high true positive rate for phishing detection while maintaining a low false positive rate for legitimate URLs. The integration of domain verification techniques, such as DNS resolution and WHOIS analysis, further improves detection reliability by identifying inactive or suspicious domains before classification.

The majority voting mechanism used in the prediction module enhances overall system performance by combining the strengths of individual models. Additionally, the inclusion of confidence scores provides transparency in predictions, allowing users to understand the certainty of each classification. The Streamlit-based interface enables real-time prediction and visualization, making the system practical and user-friendly.

Overall, the results indicate that the proposed system is effective, efficient, and suitable for real-world deployment. It not only improves phishing detection accuracy but also enhances user awareness and security by providing clear and interpretable outputs.







CONCLUSION

This paper presented a machine learning-based phishing URL detection system that effectively classifies URLs as either phishing or legitimate using algorithms such as Logistic Regression, Random Forest, and Support Vector Machine (SVM). The proposed system demonstrates high accuracy and reliable performance by leveraging multiple models and combining their predictions to improve robustness and consistency. The integration of feature extraction techniques and domain verification methods further enhances the effectiveness of the detection process.

For future work, the system can be extended by incorporating advanced deep learning models to capture more complex patterns in phishing URLs. Additionally, integrating real-time threat intelligence, developing browser extension support for live detection, and applying more sophisticated feature extraction techniques can significantly improve detection accuracy and scalability. These enhancements will make the system more adaptive, efficient, and suitable for real-world cybersecurity applications.

REFERENCES

[1] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, “Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs,” *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2009.

[2] R. Verma and N. Hossain, "Semantic Feature Selection for Text with Application to Phishing Email Detection," *IEEE Security & Privacy*, 2017.

[3] *Scikit-learn Documentation*. Available: https://scikit-learn.org

[4] *Streamlit Documentation*. Available: https://docs.streamlit.io

[5] M. Sahingoz, B. Buber, O. Demir, and B. Diri, "Machine Learning Based Phishing Detection from URLs," *Expert Systems with Applications*, 2019.

[6] *Python Documentation*, Python Software Foundation. Available: https://www.python.org

[7] *Unified Modeling Language (UML) Specification*, Object Management Group (OMG).