

Phishing Website Detection using Dimensionality Reduction and Machine Learning

Velagati Sessa Bhargavi¹, Kilim Eeshani², Panthula Sirichandana³, Reshma Begum⁴ and Souda Mythili⁵

¹⁻⁵G. Narayanamma Institute of Technology and Science/Information Technology, Hyderabad, India

Email: seshabhargavi@gnits.ac.in, eeshani1905@gmail.com, panthulasirichandana@gmail.com, reshmabegum3804@gmail.com, mythilisouda@gmail.com

Abstract— Phishing attacks have become an increasingly prevalent threat, with scammers masquerading as trustworthy entities to deceive victims into divulging sensitive information such as usernames, passwords, and credit card details. These malicious attempts have resulted in significant financial losses and compromised personal data for countless individuals. In response to this growing concern, our research aims to develop a robust anti-phishing website framework. To achieve this, we employed a range of supervised machine learning algorithms to distinguish between legitimate and phishing websites. Our approach combined the strengths of Support Vector Machine (SVM), Decision Tree, Random Forest, and XGBoost classifiers to create a hybrid detection model. By analyzing the characteristics of phishing and legitimate websites, we identified key features that enabled our model to accurately detect phishing attempts. Our evaluation of various feature selection-classification algorithms revealed that our hybrid technique achieved an impressive accuracy rate of 98.28%. This result demonstrates the effectiveness of our approach in detecting phishing websites, positioning it alongside state-of-the-art techniques in this field.

Keywords— Phishing website detection, Machine Learning, Legitimate Websites, Classification Algorithms.

I. INTRODUCTION

Phishing attacks pose a significant threat to online security, exploiting vulnerabilities in user trust to access sensitive information and financial assets. These attacks frequently disguise themselves as legitimate websites or emails, deceiving individuals into divulging personal data. While traditional approaches to detecting phishing websites have utilized supervised machine learning algorithms, there is a pressing need for more effective techniques that can adapt to evolving phishing tactics. To address this challenge, our proposed method seeks to enhance detection accuracy by leveraging a comprehensive set of features from both phishing and legitimate websites. Current phishing detection systems often rely on blacklist-based approaches and basic feature analysis, which can block or alert users to potentially harmful sites. However, these systems are limited in their ability to detect new phishing sites that are not yet included in the database. Heuristic systems analyze features such as URL structure, domain age, and page content, but may not effectively capture complex phishing tactics. Advanced systems utilizing machine learning algorithms to analyze features often rely on a fixed set of features. This can result in high computational costs and limited detection accuracy due to the sheer volume of features and data being processed.

To improve phishing website detection, our proposed system will focus on applying dimensionality reduction techniques to enhance accuracy and efficiency. This involves selecting between feature selection techniques to identify the most relevant features or applying feature extraction methods like Principal Component Analysis (PCA) to reduce dimensionality. The selected method will be used to preprocess the data, train the model, and evaluate its performance for optimal phishing detection.

By optimizing the feature set, our system aims to improve detection accuracy while reducing computational overhead. The proposed system is expected to enhance phishing detection accuracy while reducing computational overhead by leveraging dimensionality reduction techniques.

Key improvements include Feature Optimization, Machine Learning Integration, Efficient Processing.

II. LITERATURE REVIEW

The papers discussed about how different authors implemented the concept of phishing website detection in order to help the users differentiate between malicious and legitimate websites.

In paper[1],the authors explored the methodology which involves extracting lexical, host, and word vector features from URLs, reducing dimensionality with PCA, and classifying them using an SVM to detect phishing websites.the limitations include limited scope of feature selection,dependence on dataset quality.

[2] Author explored the methodology which involves preprocessing the dataset using PCA for dimensionality reduction, followed by feature selection using univariate selection techniques. The selected features are then classified using a hybrid algorithm to produce the final output.the limitations include data quality and quantityfalse positives and negatives.

In the paper[3], the author further discusses the study that evaluates various machine learning algorithms (including Random Forest and ANN) for phishing detection across three datasets, showing that these algorithms achieve over 97% accuracy in distinguishing phishing from legitimate websites.the limitations includes limited number of datasets and a lack of exploration into newer machine learning techniques.

In the paper[4], the author discussed the methodology that evaluates phishing detection accuracy by analyzing website features using various machine learning algorithms and performance metrics. The drawbacks include limited feature sets that may not capture all phishing variations and potential overfitting of machine learning models to specific datasets.

In the work [5] the author evaluates the methodology that involves combining SVM, Decision Tree, Random Forest, and XGBoost through bagging and boosting techniques to create a hybrid classifier.the drawbacks include increased computational complexity longer training times due to the combination of multiple algorithms.These studies complement each other and point the way in which the phishing and legitimate websites are differentiated with atmost accuracy and reduced computational complexity.

III. PROPOSED SYSTEM AND METHODOLOGY

A. Proposed System

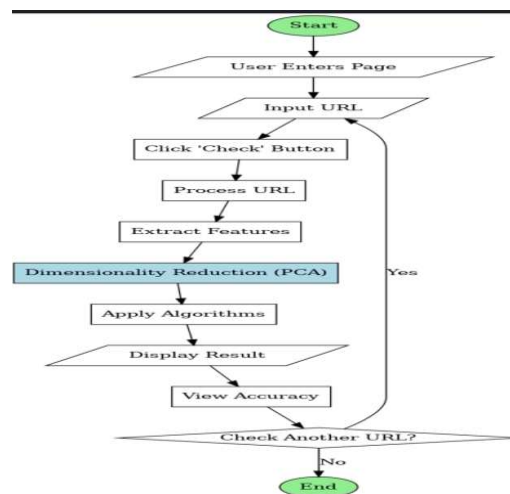


Fig. 1 System Architecture of proposed system

The proposed system employs an approach to detect phishing websites by leveraging machine learning algorithms and Principal Component Analysis (PCA) for dimensionality reduction. Upon user input of a URL, the system initiates processing by extracting a comprehensive set of features. These features encompass various indicators, including domain age, HTTPS presence, special character count, and other pertinent factors. Following feature extraction, PCA is applied to reduce the feature set, retaining only the most critical data points. This dimensionality reduction step enhances model efficiency by eliminating redundant or less impactful features. Subsequent to feature reduction, the system utilizes diverse machine learning algorithms, such as Decision Trees, Random Forest, and Support Vector Machines (SVM), to classify the URL as either phishing or legitimate. This classification enables the system to provide users with a reliable indication of the website's authenticity. The final result is displayed to the user, indicating whether the entered URL is phishing or safe to visit. Additionally, the system provides an option to view the accuracy percentages of different machine learning models used in the classification process.

B. Methodology

- Requirement Gathering and Analysis: The system functions were defined by gathering input from stakeholders, focusing on URL input, phishing detection, and algorithm performance analysis. User stories were documented to outline interactions between the system and users, detailing how users check URLs for phishing threats and view algorithm accuracy.
- System Design: The system architecture was designed, including the database schema and relationships between users, URLs, and algorithm results. Flow diagrams were created to illustrate the process, from user input to result display, including feature extraction, PCA application, and classification.
- Development: The development phase involved constructing both the front-end and back-end. Core modules, such as URL input, phishing detection, algorithm comparison, and result display, were implemented using web technologies. Additionally, feature extraction and PCA-based dimensionality reduction were integrated to enhance the performance of machine learning classifiers.
- Testing: The system was rigorously tested to ensure all functionalities worked as expected. Functional testing, unit testing, integration testing, and user acceptance testing were conducted to verify stability, accuracy, and user experience. Special attention was given to the accuracy of phishing detection algorithms and the impact of PCA on performance.
- Deployment and Maintenance: After successful testing, the system was deployed on a live server. Routine maintenance was scheduled to ensure stability, fix potential bugs, and improve algorithm efficiency. Future updates will include enhancements to phishing detection accuracy, real-time monitoring, and improved visualization of algorithm performance.

IV. RESULTS



Fig. 2 shows the interface with a textbox and CheckURL button where we as users can enter the URL that we want to check to find whether the URL is legitimate or not.



Fig. 3 shows an example where the user enters any website URL to check whether the URL is legitimate or Not.

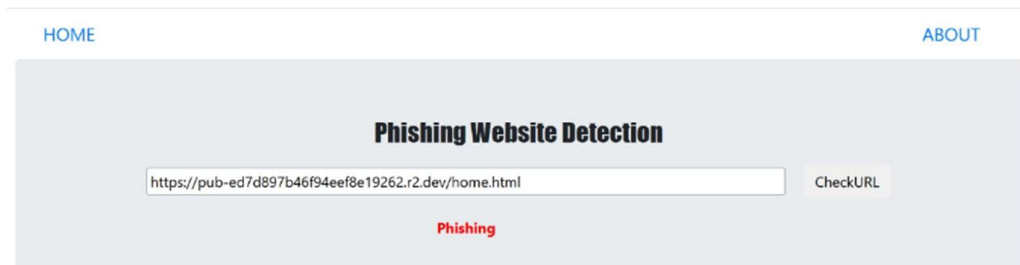


Fig. 4 shows the case where the website is evaluated to be Phishing



Fig. 5 shows the case where the website is evaluated to be Legitimate

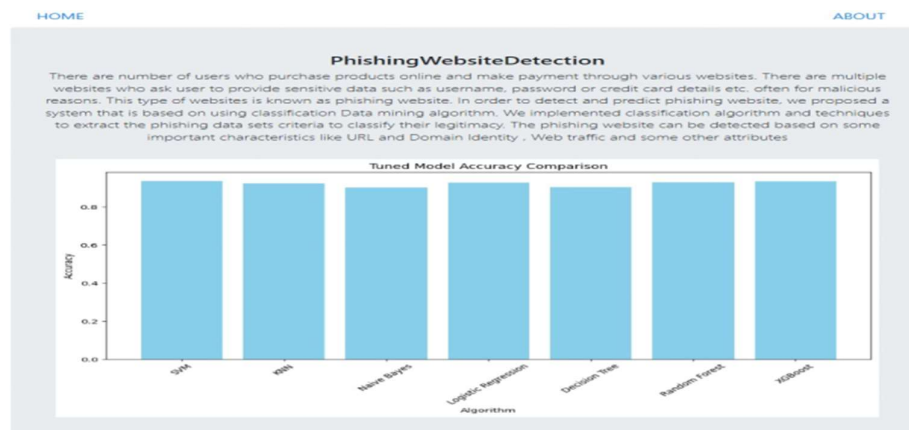


Fig. 6 explains accuracy comparison of various Algorithms applied on the mode

V. CONCLUSIONS

The proposed phishing detection system offers a robust solution to the challenges inherent in identifying phishing websites. By incorporating dimensionality reduction techniques, the system optimizes feature selection, thereby enhancing detection accuracy while minimizing computational overhead. This approach provides a more efficient alternative to traditional methods, enabling the system to adapt to evolving phishing tactics and ensuring robust cybersecurity measures. The reduced complexity of the system facilitates accelerated processing, allowing for real-time threat identification and safeguarding users against malicious attacks. The significance of this system extends beyond detection efficiency, as it contributes to the overall improvement of cybersecurity strategies by optimizing feature sets and minimizing false positives. The integration of dimensionality reduction techniques makes this system a scalable and adaptable solution for combating phishing threats in an ever-evolving digital landscape.

Future enhancements could involve integrating real-time monitoring capabilities, cloud-based analysis for broader accessibility, and the inclusion of deep learning models to further improve detection accuracy. Additionally, expanding the system to detect phishing attempts in emails and social media platforms would enhance its usability and effectiveness in protecting users from cyber threats.

REFERENCES

- [1] Junaid Rashid, Toqeer Mahmood (Corresponding Author), Muhammad Wasif Nisar. "Phishing Detection Using Machine Learning Technique," 2020.
- [2] Suhani Jain, Dr. Naveen Choudhary, Kalpana Jain. "Phishing Websites Detection Using Machine Learning," 2022.
- [3] Sohail Ahmed Khan, Wasiq Khan, Abir Hussain. "Phishing Attacks and Websites Classification Using Machine Learning and Multiple Datasets (A Comparative Analysis)," 2022.
- [4] R. Kiruthiga, D. Akila. "Phishing Detection Using Machine Learning Techniques," 2019.
- [5] Nusrath Tabassum , Farhin Faiza Neha , Md. Shohrab Hossain , and Husnu S. Narman. " A Hybrid Machine Learning based Phishing Website Detection Technique through Dimensionality Reduction" 2021. [IEEE].