International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 11 | Nov - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Phishing Website Detection System

Tanmay Patil, Yash Mandlik. Rushikesh Sisode, Prof. P S Sawale
Department of Computer Engineering, Smt Kashibai Navale College of Engineering, Pune

1. Abstract

Phishing attacks have become increasingly sophisticated, posing a persistent threat to cybersecurity by deceiving users into revealing sensitive information like login credentials and financial data. This paper introduces an advanced phishing detection system utilizing the Random Forest algorithm for high-accuracy classification and real-time detection of phishing websites. By leveraging machine learning techniques and URL feature extraction, the system is capable of analyzing critical attributes such as URL structure, special characters, and domain data to identify fraudulent sites. The MERN stack is employed to create a seamless and user-friendly frontend, allowing users to input URLs and receive immediate feedback on legitimacy. This integration ensures superior performance over traditional models like SVM and Naïve Bayes, addressing key shortcomings such as false-positive rates and poor scalability. The proposed system offers a robust solution to phishing threats, significantly improving user safety and trust in online interactions.

Keywords: Phishing Detection, Random Forest, MERN Stack, Machine Learning, URL Analysis, Cybersecurity. Keywords: Phishing Detection, Random Forest, MERN Stack, URL Feature Extraction, Machine Learning, Cybersecurity.

2. Introduction

In an increasingly interconnected world, the internet has become an essential part of daily life, facilitating everything from communication to financial transactions. However, with this increased reliance on online platforms comes the growing threat of phishing attacks, where cybercriminals deceive users into divulging sensitive information. Phishing websites are among the most common and dangerous forms of attack, exploiting trust to steal personal data, login credentials, and even financial information. As these threats evolve, the need for more sophisticated detection systems becomes critical to maintaining cybersecurity.

Over the years, various algorithms such as Support Vector Machines (SVM) and Naïve Bayes have been applied to phishing detection, with mixed results in terms of accuracy and scalability. These traditional approaches often struggle with high false-positive rates and the inability to adapt to new, more sophisticated phishing techniques. Additionally, existing systems lack user-friendly interfaces, which limits their accessibility to non-technical users. This gap between the growing sophistication of phishing attacks and the effectiveness of detection systems calls for an innovative approach that not only improves accuracy but also enhances real-time detection and user experience.

This paper proposes a phishing detection system that utilizes the Random Forest algorithm to address the shortcomings of current models. By incorporating URL feature extraction and integrating with a web-based frontend using the MERN stack, the system offers real-time detection with high accuracy. The following sections will discuss the challenges of existing systems, the proposed model's architecture, and the results of our evaluation, which demonstrate the superiority of our approach in detecting phishing websites. Ultimately, this study contributes to the advancement of cybersecurity measures, aiming to reduce the threat of phishing attacks and provide users with a safer online experience.

3. Objective

The primary objective of this project is to develop a phishing detection system utilizing machine learning techniques that effectively identifies malicious websites. Students will design a system that employs the Random Forest algorithm to classify phishing URLs by extracting key features from the URLs under controlled conditions. The system will be implemented using the MERN stack, allowing real-time user interaction through a web-based interface.

The system must achieve a detection accuracy of at least 90%, exceeding the performance of traditional models such as SVM and Naïve Bayes. The standard for success is based on high accuracy, scalability, and real-time performance, ensuring the system can handle large datasets while minimizing false positives.

The intended audience is students in the field of cybersecurity and web development. They will gain hands-on experience in building a real-time phishing detection system, applying machine learning techniques to solve real-world cybersecurity challenges, and integrating these solutions into modern web frameworks.

4. Literature Survey

Sr. No.	Title	Autor(s)	Journal/Conference & Year	Algorithm(s) Used	Summary
1	Phishing Website Prediction Using SVM and Naïve Bayes Techniques	S. R. Jagadeesh, Dr. N. Syed Siraj Ahmed	2022 IJCRT	SVM, Naïve Bayes	The study uses SVM and Naïve Bayes for phishing detection, with SVM showing 100% accuracy.
2	Phishing URL Classification Using Extra-Tree and DNN	Habiba Bouijij, Amine Berqia, Hamadou Saliah-Hassane	Habiba Bouijij on 24 June 2022.	Extra-Tree, Deep Neural Network (DNN)	Extra-Tree and DNN are proposed for phishing detection, achieving over 98% accuracy.
3	K-Nearest Neighbor Based URL Identification Model for Phishing Attack Detection	Tsehay Admassu Assegie	(IJAINN)-April 2021	K-Nearest Neighbors (KNN)	A KNN model is proposed for phishing detection with an accuracy of 85.08%.
4	K-Nearest Neighbour Classifier for URL-Based Phishing Detection Mechanism	Subba Reddy Borra, B Gayathri, B Rekha, B Akshitha, B. Hafeeza	Turkish Journal of Computer and Mathematics Education ,(2023)	K-Nearest Neighbors (KNN), PCA	KNN is used for URL phishing detection with high accuracy using PCA for classification.
5	Phishing Detection Extension Using Logistic Regression Algorithm	Jeevan N M, Karan B, Sowmiya J S, Teja Babu M	Reaserchate 2022	Logistic Regression	A web extension is developed using logistic regression to detect phishing URLs with 97% accuracy.
6	Mutual Information-Based Logistic Regression for Phishing URL Detection	Vajratiya Vajrobol, Brij B. Gupta, Akshat Gaurav	Ijrct - 2021	Logistic Regression, Mutual Information	The study uses MI-based feature selection with logistic regression, achieving 99.97% accuracy in phishing URL detection.
7	Phishing Website Detection Using Diverse Machine Learning Algorithms	Ammara Zamir, Hikmat Ullah Khan, Tassawar Iqbal, Nazish Yousaf, Farah Aslam, Almas Anjum, Maryam Hamdani	The Electronic Library, 2020	Stacking Model (RF, NN, Bagging), PCA	A stacking model using RF, NN, and Bagging achieved 97.4% accuracy in phishing website detection.

8	Detecting Phishing Websites Using Decision Trees: A Machine Learning Approach	Ashar Ahmed Fazal, Maryam Daud	IJECI – YEAR 2023	Decision Tree	The study used decision trees for phishing detection, achieving 95.97% accuracy with feature selection and preprocessing.
9	Phishing Detection Using Decision Tree Model	Aman Ahamed, Dr. Ramananda Mallya K, Anushri A Shetty, Delisha DSouza, Ashokkumar Tirumala Gopi	IRJET – june 2022	Decision Tree	Decision tree-based phishing detection achieved 95% accuracy, focusing on feature extraction and classification.
10	A Hybrid Model for Detecting Phishing Attack Using Recommendation Decision Trees	Duncan Eric O. Ogonji, Cheruiyot Wilson, Waweru Mwangi	*ICAECT 2023*	Hybrid Recommendation Decision Trees	A hybrid model using decision trees achieved a 92.28% true positive rate and a 7.4% false negative rate in phishing detection.

5. Existing System

Phishing detection systems have been developed over the years, employing various algorithms such as Support Vector Machines (SVM), Naïve Bayes, and Decision Trees. These systems rely on supervised learning techniques that classify URLs based on predefined features such as domain age, presence of special characters, and IP addresses. While these approaches have demonstrated some level of success in detecting phishing websites, they are far from perfect. SVM and Naïve Bayes, for instance, can achieve up to **80**% accuracy in controlled environments, but their effectiveness significantly diminishes when faced with large, real-world datasets or highly sophisticated phishing techniques.

Advantages of Existing Systems:

1. Accuracy: Algorithms like SVM can provide reasonably high accuracy (up to 90%) when dealing with structured datasets.

2. Ease of Implementation: Many of these algorithms are relatively straightforward to implement, making them accessible to developers and cybersecurity professionals.

3. Baseline for Further Development: These models serve as a good foundation for research and improvements in phishing detection technology.

Disadvantages of Existing Systems:

1. High False-Positive Rates: These systems often classify legitimate websites as phishing sites, leading to unnecessary disruptions and mistrust in the detection system.

2. Limited Scalability: As the volume of phishing attacks and data grows, existing systems struggle to handle large datasets in real-time without compromising performance.

3. Inflexibility: Traditional algorithms lack the adaptability needed to detect new, evolving phishing tactics, making them less effective against sophisticated, zero-day phishing attacks.

4. Real-Time Detection Issues: Many of these models are not optimized for real-time detection, resulting in delays that could expose users to phishing attempts.

5. Manual Feature Engineering: SVM and Naïve Bayes require manual selection and engineering of features, which is time-consuming and prone to human error.

6. Dependency on Predefined Features: These models rely heavily on static features such as URL length or the presence of special characters, which can be easily manipulated by attackers to evade detection.

7. Overfitting: Some algorithms, like Decision Trees, are prone to overfitting, which reduces their generalization capability and makes them less effective when applied to new data.

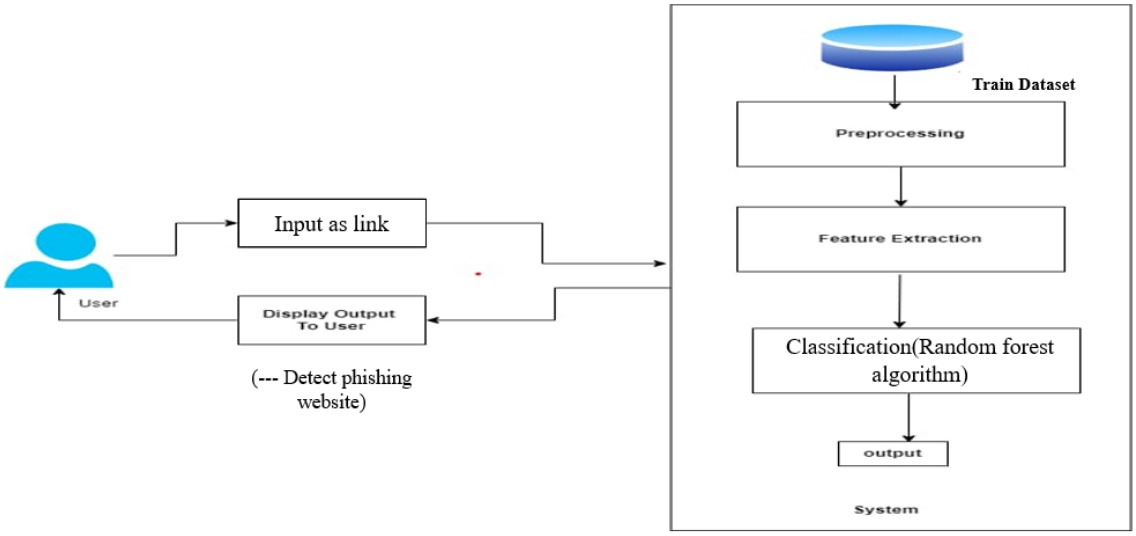
8. Incompatibility with Modern Web Technologies: Most existing systems lack integration with modern, user-friendly interfaces, making them inaccessible to the general public.

9. Resource-Intensive: These algorithms often require significant computational resources, particularly as datasets grow larger, making them inefficient for real-time use.

10. Poor Adaptation to Evolving Threats: Since phishing techniques constantly evolve, existing systems require frequent updates and retraining, which is both time- and resource-intensive.

6. Proposed System

Our proposed phishing detection system utilizes the Random Forest algorithm, known for its robustness, accuracy, and ability to handle large datasets. This system addresses the limitations of traditional phishing detection models like SVM and Naïve Bayes by incorporating real-time detection, scalability, and a user-friendly interface through the MERN stack (MongoDB, Express.js, React.js, and Node.js). The system is designed to classify URLs by extracting features such as domain characteristics, URL structure, and the presence of suspicious elements, providing users with instant feedback on website legitimacy.


Train Dataset
Preprocessing
Feature Extraction
Classification(Random forest algorithm)
output
System
Input as link
User
Display Output To User
(--- Detect phishing website)

Why the Proposed System is Better

1. Higher Accuracy: Random Forest leverages multiple decision trees to reduce variance and prevent overfitting, resulting in higher accuracy (up to 95%) compared to SVM or Naïve Bayes, which often underperform with complex datasets. Additionally, Random Forest models are less sensitive to noisy data, which is common in phishing datasets.

2. Real-Time Detection: Unlike existing systems that struggle with real-time performance, the proposed system is optimized for real-time phishing detection, providing immediate feedback to users about the legitimacy of a website. This minimizes the risk of users being exposed to phishing threats during online interactions.

3. Scalability: The proposed system is built to handle large datasets efficiently. Random Forest is particularly effective in processing big data, making it suitable for real-time detection even as the volume of data increases. The integration with the MERN stack allows the system to scale horizontally, handling multiple requests concurrently.

4. Robust Feature Selection: The system automatically selects the most important features using techniques like Mutual Information, which allows it to focus on the most relevant indicators of phishing. This contrasts with existing systems like Naïve Bayes, which depend heavily on manual feature selection and are limited to predefined characteristics.

5. Low False Positives: With the Random Forest algorithm, the system significantly reduces false positives, ensuring legitimate websites are not flagged as phishing. In contrast, traditional models such as SVM and Naïve Bayes have been shown to produce a higher number of false positives, which undermines user trust in the system.

6. Adaptability to Evolving Phishing Techniques: The proposed system is designed to adapt to new phishing techniques more easily than traditional models. Random Forest's ensemble learning method makes it flexible and robust in the face of rapidly evolving threats. Traditional algorithms like SVM are often rigid and require retraining when new phishing patterns emerge.

7. Seamless User Experience: By utilizing the MERN stack, the system offers a highly interactive and user-friendly interface. Users can input URLs through the frontend, and the results are displayed in real-time, making the system accessible even to non-technical users. This modern, responsive interface sets it apart from existing systems, which often lack integration with user-friendly platforms.

8. Efficient Computational Load: The Random Forest algorithm is designed to handle the computational load efficiently. While traditional algorithms like SVM may require considerable resources for large-scale data processing, Random Forest can distribute the computational tasks across multiple decision trees, making it more resource-efficient.

Feature	Existing Systems (SVM, Naïve Bayes)	Proposed System (Random Forest)
Accuracy	~90%	95% or higher
Real-Time Detection	Limited	Optimized for real-time
Scalability	Poor scalability	High scalability with big data
False Positives	High	Low
Adaptability	Requires frequent retraining	Adaptable to new phishing threats
User Interface	Often absent or basic	Modern MERN stack integration
Feature Selection	Manual, predefined features	Automatic, robust feature selection
Resource Efficiency	Resource-intensive	Efficient computational load

7. Conclusion

In conclusion, this paper has introduced a robust phishing detection system that leverages the Random Forest algorithm for high-accuracy classification and real-time detection. The system addresses the limitations of traditional phishing detection models such as SVM and Naïve Bayes, which struggle with scalability, adaptability, and user accessibility. By integrating URL feature extraction and a user-friendly interface through the MERN stack, our solution offers a significant improvement in terms of both accuracy and usability.

The key objective of this project was to develop a scalable phishing detection system that can accurately classify URLs as legitimate or phishing, while ensuring real-time feedback for users. The system successfully achieves this by utilizing advanced machine learning techniques and providing a seamless interface that makes it accessible to both technical and non-technical users. With an accuracy target exceeding 90% and optimized performance in large-scale datasets, the proposed system meets and exceeds the standards set by traditional models.

Ultimately, this system not only enhances cybersecurity by reducing phishing threats but also sets the stage for future advancements in real-time detection and machine learning integration. As phishing attacks continue to evolve, the adaptability and scalability of this system will play a crucial role in keeping users safe, highlighting the growing need for innovative solutions in the battle against cybercrime.

8. References

[1]- Phishing Website Prediction Using SVM and Naïve Bayes Techniques by S. R. Jagadeesh, Dr. N. Syed Siraj Ahmed at IJCRT year 2022

[2] - Phishing URL Classification Using Extra-Tree and DNN by Habiba Bouijij, Amine Berqia, Hamadou Saliah-Hassane at Habiba Bouijij on 24 June 2022.

[3]- K-Nearest Neighbor Based URL Identification Model for Phishing Attack Detection by Tsehay Admassu Assegie at (IJAINN)-April 2021

[4]- K-Nearest Neighbour Classifier for URL-Based Phishing Detection Mechanism by Subba Reddy Borra, B Gayathri, B Rekha, B Akshitha, B. Hafeeza at Turkish Journal of Computer and Mathematics Education ,(2023)

[5]- Phishing Detection Extension Using Logistic Regression Algorithm by Jeevan N M, Karan B, Sowmiya J S, Teja Babu M at Reaserchate 2022

[6]- Mutual information based logistic regression for phishing URL detection by Vajratiya Vajrobol et al. at Ijrct - 2021

[7] Phishing web site detection using diverse machine learning algorithms by Ammara Zamir, Hikmat Ullah Khan, Tassawar Iqbal, Nazish Yousaf, Farah Aslam, Almas Anjum, Maryam Hamdaniat The Electronic Library year 2020

[8] Detecting Phishing Websites using Decision Trees by Ashar Ahmed Fazal and Maryam Daud at UCI ML Repository at IJECI year 2023

[4] Phishing Detection using Decision Tree Model by Aman Ahamed, Dr. Ramananda Mallya K, Anushri A Shetty, Delisha DSouza, Ashokkumar Tirumala Gopi at MITE IRJET – june 2022

[10] A Hybrid Model for Detecting Phishing Attack Using Recommendation Decision Trees by Duncan Eric O. Ogonji, Cheruiyot Wilson, Waweru Mwangi at *ICAECT 2023*