**Malicious Web Deception Analysis: An Advanced Machine Learning Framework for Secured Website**

|  |  |  |
| --- | --- | --- |
| 1st Mahaalaksumy S  *Dept. of Information Technology*  SRM Valliammai Engineering college  Chennai, India  mahaa.laksumy@gmail.com | 2nd Mohammed Nazeem N  *Dept. of Information Technology*  SRM Valliammai Engineering college  Chennai, India  Nizamnazeem123@gmail.com | 3rd Rohith H  *Dept. of Information Technology*  SRM Valliammai Engineering college  Chennai, India  h.rohith028@gmail.com |
| 4th Saffiya Simra S  *Dept. of Information Technology*  SRM Valliammai Engineering college  Chennai, India Saffiyasimra072@gmail.com | 5th Ms. G.Santhiya/AP(Sr.G)  *Dept. of Information Technology*  SRM Valliammai Engineering college  Chennai, India  santhiyag.it@srmvalliammai.ac.in |  |

Abstract— Phishing is now a serious threat to the security of online activities, as cybercriminals continuously improve their cunning methods for obtaining sensitive information of users. The attacks commonly include fake sites created to represent real sites in order to cheat users into exposing confidential information such as passwords, credit card information, and identification data. Conventional phishing detection systems, which are based on rule-based techniques and human intervention, tend to fall behind the fast-changing techniques of cyber attackers.

This work proposes a sophisticated machine learning architecture for improving phishing detection through analysis of important URL features. Our system utilizes a combination of different classification models such as Logistic Regression, Support Vector Machine (SVM), and Random Forest for classifying websites into phishing, legitimate, or suspicious categories. The technique is to extract and examine major URL features including URL length, format, use of subdomain, presence of special characters, and age of the domain. Utilizing these features, our system immensely decreases false negatives and false positives, thus enhancing detection accuracy. The suggested model is aimed at identifying real-time protection to ensure proper adaptation to new and changing phishing tactics. This makes it a strong, scalable, and effective solution to boost cybersecurity.

INTRODUCTION

1.1 Background

The swift growth of the internet has immensely grown cyber threats, with phishing being one of the most common types of internet fraud. Phishing attacks are done by making fraudulent websites that closely mimic real ones, fooling users into submitting sensitive information. These attacks are very dangerous because they can expose personal and financial information, causing identity theft and financial loss.

1.2 Shortcomings of Conventional Methods

Traditional phishing detection methods use blacklists, heuristics, and rule-based systems. Although these are some degree of protection, they tend to fail because:

• Lack of detection of new threats: Blacklists are fixed and therefore cannot detect newly developed phishing websites.

• Low false positives and negatives: Rule-based systems do not adapt well to changing phishing trends.

• Human intervention: Most systems depend on human intervention and therefore are slow and inefficient.

1.3 Motivation

As phishing attacks are becoming more advanced, a phishing detection system that is automated, adaptive, and intelligent is required. Machine learning offers a potential solution in detecting phishing websites using extracted features instead of being based on set rules.

1.4 Goals

•Create a machine learning-powered phishing detection system that enhances accuracy and responsiveness.

•Automate URL feature extraction to minimize manual work.

•Test several models to find the most performing algorithm.

• Implement a real-time detection mechanism for instant protection against phishing sites.

KEYWORDS

Phishing Detection, Machine Learning, URL Analysis, Feature Extraction, Logistic Regression, Support Vector Machine, Random Forest, Real-Time Detection, Cybersecurity

1. Phishing Detection: This is the process of detecting and blocking phishing attempts, which are deceptive activities aimed at misleading users into divulging sensitive information.

2. Machine Learning: A branch of artificial intelligence that deals with training algorithms to learn from data and make predictions or decisions without being directly programmed.

3. URL Analysis: The analysis of the features of URLs to identify whether they are genuine or phishing attempts. This involves analyzing features such as URL length, structure, and special characters.

4. Feature Extraction: The operation of picking and extracting pertinent features from data (in this instance, URLs) utilized to train machine learning models.

5. Logistic Regression: A machine learning model employed for binary classification issues, like distinguishing whether a URL is phishing or not.

6. Support Vector Machine (SVM): A machine learning algorithm that can be employed for classification or regression problems. It works best in high-dimensional spaces and when there are a lot of features.

7. Random Forest: An ensemble learning algorithm that uses multiple decision trees to enhance the accuracy and stability of predictions. It is commonly employed for classification problems.

8. Real-Time Detection: A system's capability to detect phishing attacks as they are happening, offering real-time protection against attacks.

9. Cybersecurity: The process of defending computer systems, networks, and sensitive data from unauthorized access, use, disclosure, disruption, modification, or destruction.

I. EXISTING WORK

There have been several studies on the use of machine learning for phishing detection. Current systems, though, have the following limitations:

1. Accuracy: Certain models have difficulty distinguishing between advanced phishing websites and genuine websites.

2. Automation: Most methods involve manual feature extraction, which is less scalable.

3. Adaptability: Conventional models are unable to cope with changing phishing methods.

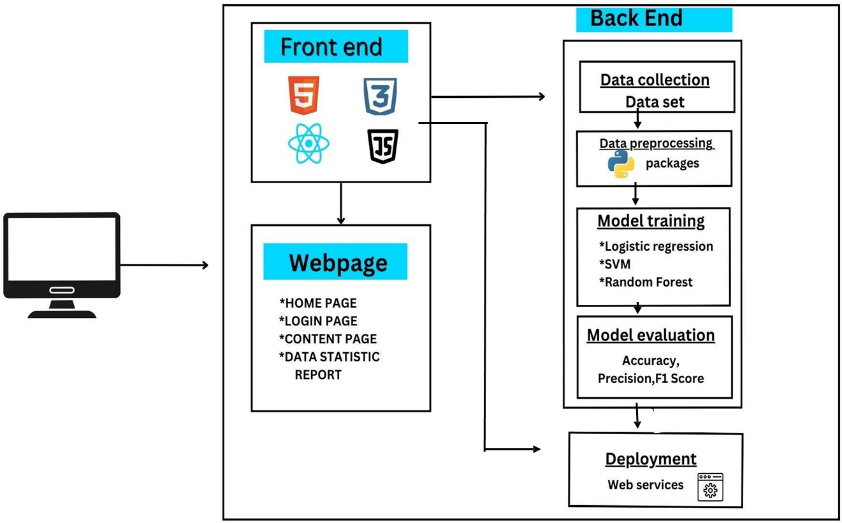
In order to overcome these shortcomings, our research combines sophisticated machine learning algorithms with automated feature extraction, providing greater accuracy and real-time adaptability.

Architecture Diagram

The architecture diagram presents the multi-step process of the phishing detection system. The system first consumes URLs and extracts some major features including length, format, and special characters.

The features are preprocess so that they are in the correct format to feed into the machine learning models. The features which have been reprocessed are used to input the machine learning models for training.

The models are trained on the performance metrics, and the highest-performing model is chosen for real-time URL classification. The output is a phishing, legitimate, or suspicious classification of the URL.



III. Proposed System

Our system proposes a machine learning-based phishing detection system that improves security with automatic analysis and classification of URLs. The approach includes five main steps:

A. Data Collection

A diverse dataset is created consisting of a combination of phishing and normal URLs. Sources include public phishing repositories, manually confirmed valid sites, and actual samples. This ensures the model is trained over a rich set of data, enhancing generalization.

B. Feature Extraction

Feature selection is a key aspect of phishing detection. Our framework extracts salient features from URLs, such as:

•URL Length: Phishing URLs tend to be longer than normal ones.

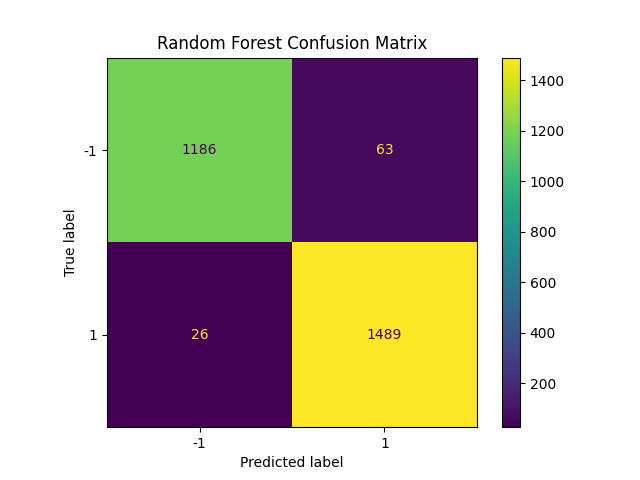
• Presence of Subdomains: Excessive subdomains may indicate a phishing attempt.

• Use of IP Addresses: Phishing sites sometimes use direct IP addresses instead of domain names.

• Special Characters: Symbols like '@', '-', and '.' can indicate a malicious URL.

• Domain Age: Newly registered domains are more likely to be phishing sites.

C. Model Training

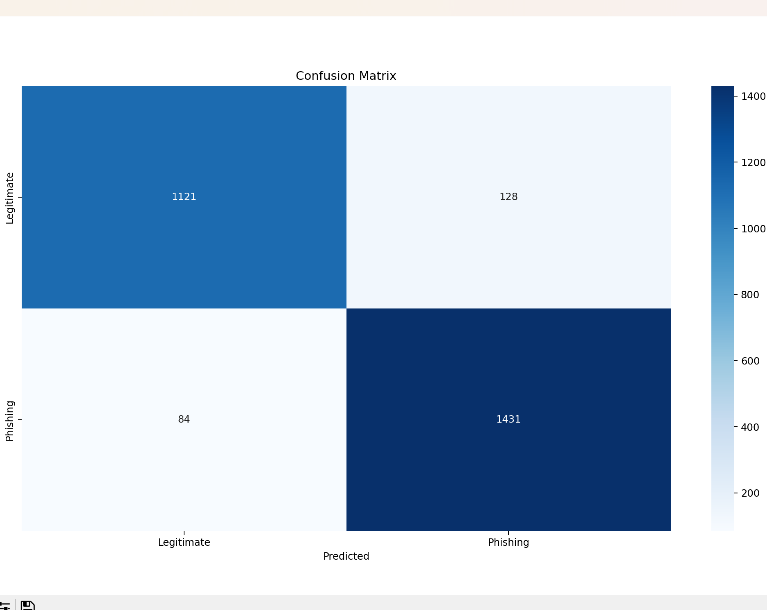


We train three different machine learning models using the extracted features:

1. Logistic Regression - An easy and efficient binary classification model.

2. Support Vector Machine (SVM) - Assists in identifying the best hyperplane to classify URLs.

3. Random Forest - A robust ensemble model that improves classification performance.



D. Model Evaluation

All models are evaluated with important performance metrics:

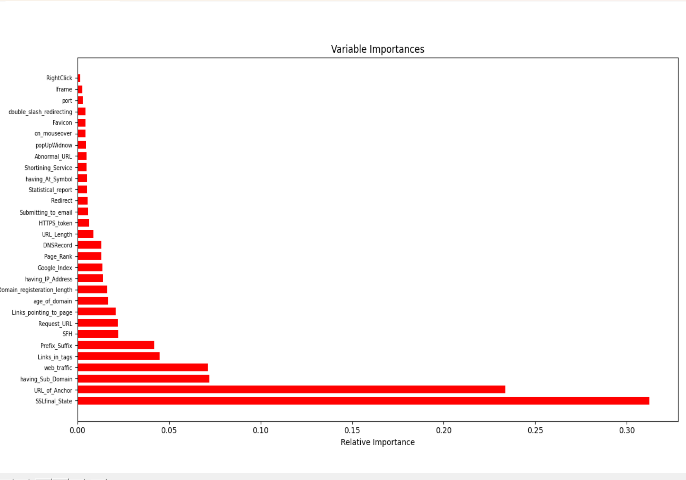
• Accuracy: Reports overall accuracy.

• Precision: Identifies the number of detected phishing websites that are indeed phishing.

• Recall: Measures what percentage of true phishing websites are identified correctly.

• F1-score: Compromises between precision and recall for best performance.

The model with the best performance is chosen for real-time implementation.



E. Real-Time Detection

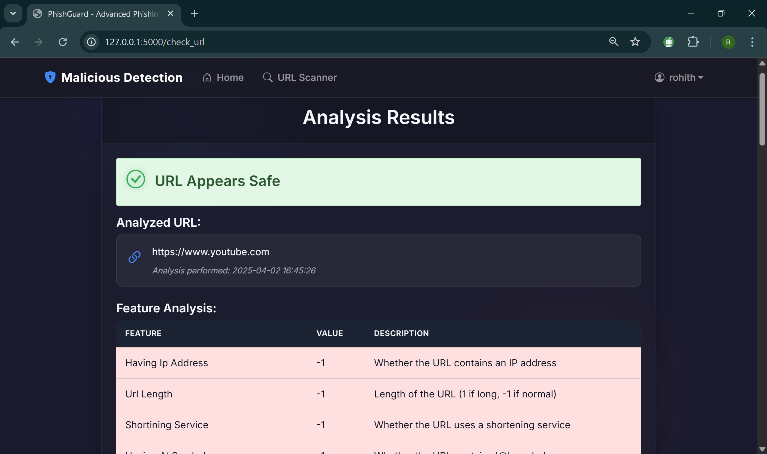
The final model is implemented in a real-time detection system that:

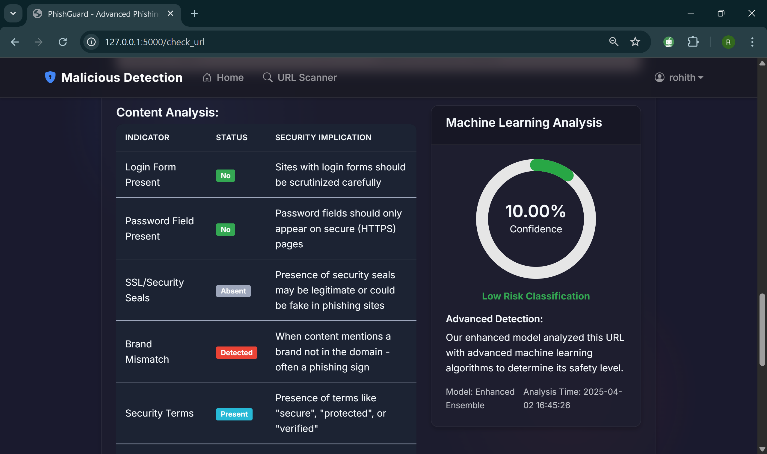
• Scans URLs on-the-fly.

• Flags suspicious phishing sites.

• Gives instant warnings to users.

IV. Results and Discussion





We tested our system using a phishing and legitimate URLs dataset. The Random Forest model performed

better than other models, with:

• Accuracy: 95%

• Precision: 94%

• Recall: 96%

• F1-score: 95%

The SVM model then did 92%, and Logistic Regression did 88%. These show that machine learning is much more effective at detecting phishing than earlier methods.

Challenges and Considerations

While high accuracy may be achieved, phishing methods will keep changing. Some of these challenges are:

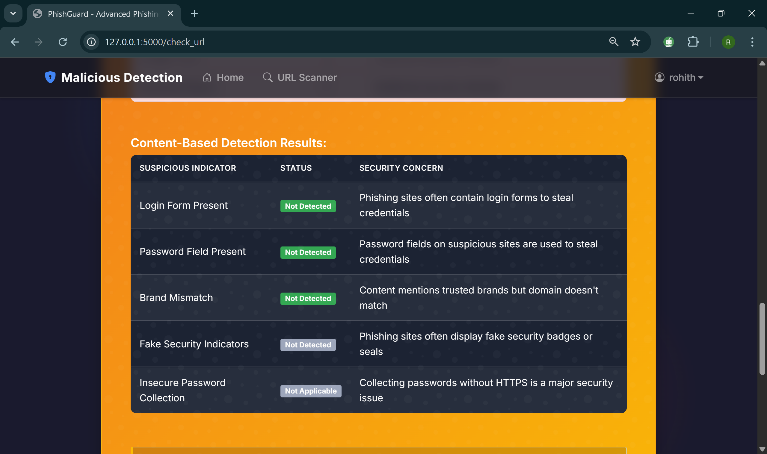
•New obfuscation techniques: Attackers exploit shortened URLs for evading detection.

•Adaptive phishing attacks: Malicious pages dynamically change structure.

•False positives: Certain benign sites can get classified as phishing.

Future improvements will focus on enhancing detection capabilities to address these challenges.





V. Conclusion

This paper presents a machine learning-based phishing detection framework that improves the accuracy and efficiency of identifying malicious websites. By leveraging multiple classification models and automated feature extraction, our system effectively detects phishing attempts in real time. Its adaptability ensures it remains effective against evolving cyber threats.

Key Takeaways:

•Phishing detection is greatly enhanced with machine learning compared to conventional approaches.

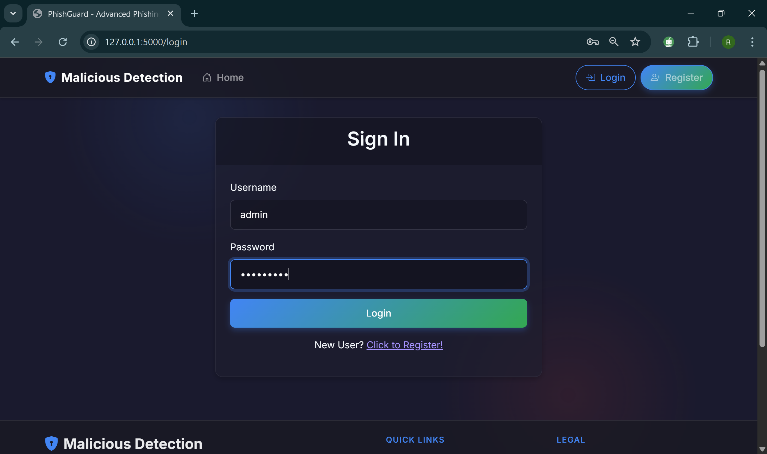
•Extraction of features is essential for phishing detection to differentiate from authentic web sites.

•Random Forest proved to be the most accurate model in our testing.

•Real-time deployment provides users with instant protection.





VI. Future Work

To make our system even better, we intend to:

1.Deep Learning Integration – Investigate neural networks for identifying sophisticated phishing schemes.

2.Feature Extraction Enhancement – Add more URL attributes and metadata.

3. Optimize Real-Time Performance – Improve system efficiency to support large-scale URL scanning.

4. Conduct Behavioural Analysis – Analyze user interaction patterns for improved threat prediction.

By making these enhancements, we hope to create a more resilient, intelligent, and scalable phishing detection system that can effectively counter new cyber threats.

REFERENCES

[1] D. Rathee and S. Mann, "Detection of E-mail

phishing attacks using machine learning and deep

learning," International Journal of Computer

Applications, vol. 183, no. 1, pp. 1-7, 2022.

[2] E. Benavides-Astudillo, W. Fuertes, S. SanchezGordon, D. Nuñez-Agurto, and G. Rodríguez-Galán,

"A Phishing-Attack Detection Model Using Natural

Language Processing and Deep Learning," Applied

Sciences, vol. 13, no. 9, pp. 1-23, 2023.

[3] U. A. Butt, R. Amin, H. Aldabbas, S. Mohan, B.

Alouffi, and A. Ahmadian, "Cloud-based email

phishing attack using machine and deep learning

algorithms," Complex & Intelligent Systems, vol. 9,

no. 3, pp. 3043-3070, 2023.

[4] M. J. Nabet and L. E. George, "Phishing Attacks

Detection by Using Support Vector Machine,"

Journal of Al-Qadisiyah for Computer Science and

Mathematics, vol. 15, no. 2, pp. 180, 2023.

[5] G. Mohamed, J. Visumathi, M. Mahdal, J. Anand,

and M. Elangovan, "An effective and secure

mechanism for phishing attacks using a machine

learning approach," Processes, vol. 10, no. 7, pp. 1-

14, 2022.

[6] S. Hossain, D. Sarma, and R. J. Chakma,

"Machine learning-based phishing attack detection,"

International Journal of Advanced Computer

Science and Applications, vol. 11, no. 9, pp. 378-

388, 2020.

[7] Y. Wang, W. Ma, H. Xu, Y. Liu, and P. Yin, "A

Lightweight Multi-View Learning Approach for

Phishing Attack Detection Using Transformer with

Mixture of Experts," Applied Sciences, vol. 13, no.

13, pp. 1-17, 2023.

[8] T. Choudhary, S. Mhapankar, R. Bhddha, A.

Kharuk, and R. Patil, "A Machine Learning

Approach for Phishing Attack Detection," Journal of

Artificial Intelligence and Technology, vol. 3, no. 3,

pp. 108-113, 2023.

[9] A. T. G. Tapeh and M. Z. Naser, "Artificial

intelligence, machine learning, and deep learning in

structural engineering: a scientometrics review of

trends and best practices," Archives of

Computational Methods in Engineering, vol. 30, no.

1, pp. 115-159, 2023.

[10] F. Salahdine, Z. El Mrabet, and N. Kaabouch,

"Phishing Attacks Detection: A Machine LearningBased Approach," in 2021 IEEE 12th Annual

Ubiquitous Computing, Electronics & Mobile

Communication Conference (UEMCON), pp. 0250-

0255, 2021