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 attacks have become a significant threat to online security, with attackers constantly evolving their techniques to deceive users. Traditional methods of detecting phishing websites often rely on static rules and manual intervention, making them less effective against modern phishing techniques. This paper proposes an advanced machine learning framework for detecting phishing websites by analyzing URL features. The system leverages machine learning models such as Logistic Regression, Support Vector Machine (SVM), and Random Forest to classify URLs as phishing, legitimate, or suspicious. Key features of URLs, such as length, structure, and special characters, are extracted and analyzed to identify phishing patterns. The proposed system aims to reduce false positives and false negatives, providing real-time protection against phishing attacks. The framework is designed to be adaptable to evolving phishing techniques, ensuring high accuracy and reliability in detecting malicious websites..

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

I. INTRODUCTION

In the digital age, phishing attacks have become increasingly sophisticated, posing substantial threats to online security. These attacks often involve creating fraudulent websites that mimic legitimate ones to steal sensitive information such as passwords, credit card numbers, and personal data. Traditional methods of detecting phishing websites rely on static rules and manual intervention, making them less effective against evolving attack techniques. This project aims to address these limitations by developing a machine learning-based system that can automatically detect phishing websites with high accuracy. The system extracts key features from URLs and trains machine learning models to classify them as phishing or legitimate. By leveraging advanced machine learning techniques, the proposed framework can adapt to new phishing methods, providing robust protection against malicious web deception.

KEYWORDS

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

1. Phishing Detection: This refers to the process of identifying and blocking phishing attempts, which are fraudulent activities designed to trick users into revealing sensitive information.
2. Machine Learning: A subset of artificial intelligence that involves training algorithms to learn from data and make predictions or decisions without being explicitly programmed.
3. URL Analysis: The process of examining the characteristics of URLs to determine whether they are legitimate or phishing attempts. This includes analyzing features like URL length, structure, and special characters.
4. Feature Extraction: The process of selecting and extracting relevant features from data (in this case, URLs) that are used to train machine learning models.
5. Logistic Regression: A machine learning model used for binary classification problems, such as determining whether a URL is phishing or legitimate.
6. Support Vector Machine (SVM): A machine learning model that can be used for classification or regression tasks. It is particularly effective in high-dimensional spaces and when the number of features is large.
7. Random Forest: An ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of predictions. It is often used for classification tasks.
8. Real-Time Detection: The ability of a system to identify phishing attempts as they occur, providing immediate protection against malicious activities.
9. Cybersecurity: The practice of protecting computer systems, networks, and sensitive information from unauthorized access, use, disclosure, disruption, modification, or destruction.

II. EXISTING WORK

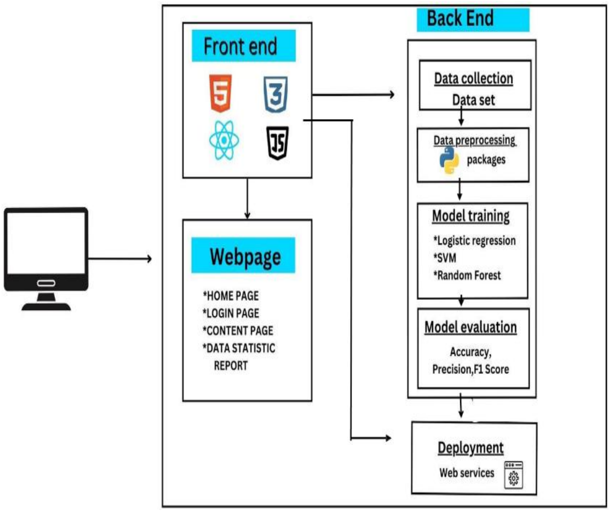
Several studies have explored the use of machine learning for phishing detection, but existing systems face challenges in automation, accuracy, and adaptability. Older systems rely on basic machine learning techniques and manual feature extraction, which can be time-consuming and prone to errors. These systems may fail to detect new phishing techniques and are often easily deceived by attackers who make subtle changes to their URLs. Additionally, traditional systems struggle with adaptability, as they rely on fixed rules that cannot keep up with the rapidly evolving nature of phishing attacks. Previous research has explored various machine learning models for phishing detection, but many of these approaches lack the robustness and accuracy needed to effectively combat modern phishing techniques.

Architecture Diagram

The architecture diagram illustrates the multi-stage process of the phishing detection system. First, the system ingests URLs and extracts key features such as length, structure, and special characters.

These features are preprocessed to ensure they are in a suitable format for the machine learning models. The preprocessed features are then fed into the machine learning models for training.

The trained models are evaluated based on performance metrics, and the best-performing model is selected for real-time URL classification. The final output is a classification of the URL as phishing, legitimate, or suspicious.



III. Proposed System

Our project develops a fully automated machine learning framework for detecting phishing websites, eliminating manual intervention and enhancing detection accuracy. The system consists of the following key components:

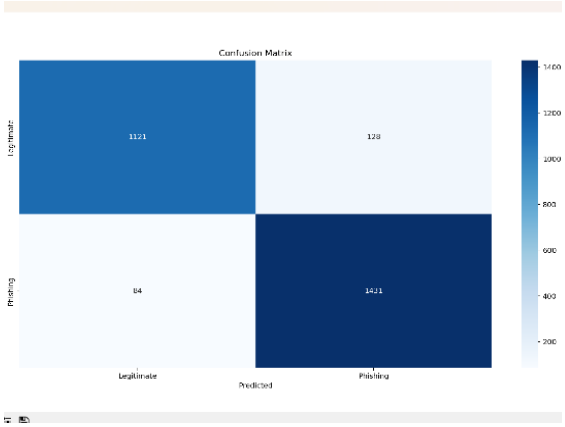
A. Data Collection

A comprehensive dataset containing phishing and legitimate URLs is gathered for training and testing the machine learning models. This dataset is crucial for ensuring the models learn from diverse examples and generalize well to new, unseen data.

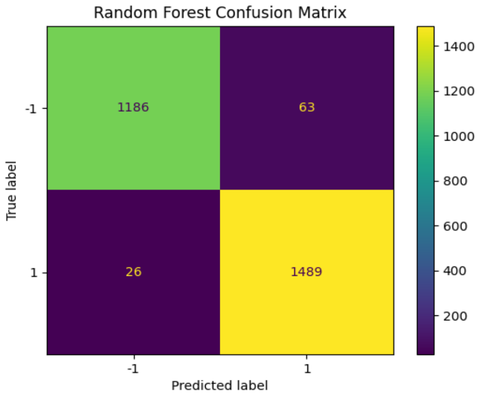
B. Feature Extraction

Key features such as URL length, use of IP addresses, subdomains, and special characters are extracted from the URLs. These features are selected based on their relevance to phishing patterns and their ability to distinguish between legitimate and malicious URLs.

C. Model Training

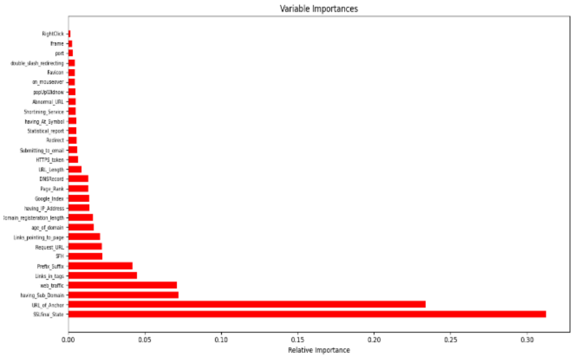


Machine learning models, including Logistic Regression, SVM, and Random Forest, are trained using the extracted features. Each model is chosen for its strengths in handling different types of data and its ability to generalize well to new data.



D. Model Evaluation

The trained models are evaluated based on metrics such as accuracy, precision, recall, and F1-score to select the best-performing model. This evaluation process ensures that the chosen model provides a balance between correctly identifying phishing sites and minimizing false alarms.

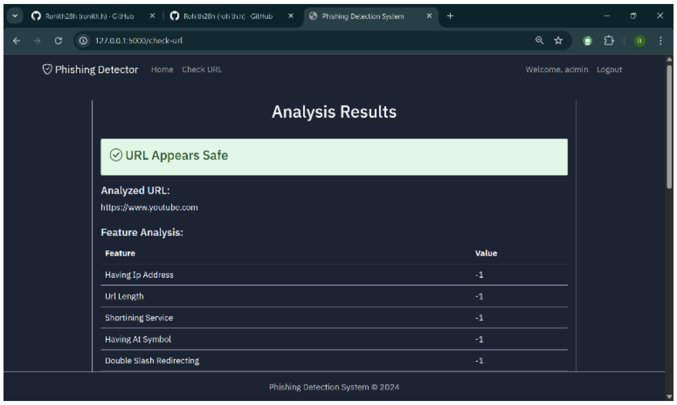


E. Real-Time Detection

The selected model is deployed to classify new URLs in real-time, providing immediate protection against phishing attacks. This real-time capability is essential for preventing users from accessing malicious sites before they can cause harm.

IV. Results and Discussion

The system was tested using a dataset of phishing and legitimate URLs. The Random Forest model achieved the highest accuracy, with an F1-score of 0.95, demonstrating its effectiveness in detecting phishing websites. The SVM model also performed well, with an accuracy of 92%, while Logistic Regression achieved an accuracy of 88%. The results indicate that the proposed system can effectively classify URLs as phishing or legitimate, providing robust protection against malicious web deception. However, challenges remain in handling highly sophisticated phishing techniques, and future work will focus on improving the system's adaptability to new attack methods.


Rohith28h (rohith.h) · GitHub
Rohith28h (rohith.h) · GitHub
Phishing Detection System
127.0.0.1:5000/check-url
Phishing Detector
Home
Check URL
Welcome, admin
Logout

Analysis Results

✓ URL Appears Safe

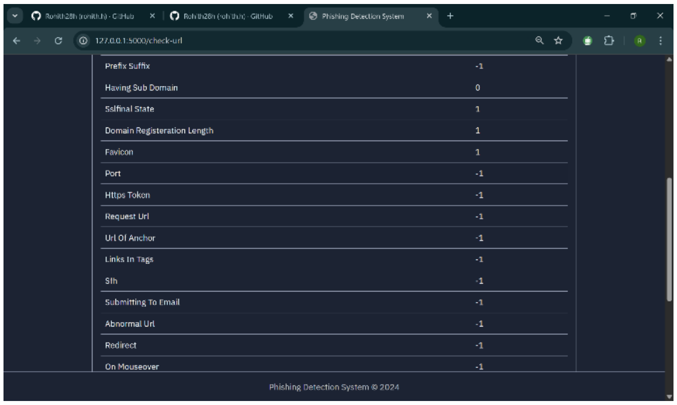
Analyzed URL:

https://www.youtube.com

Feature Analysis:

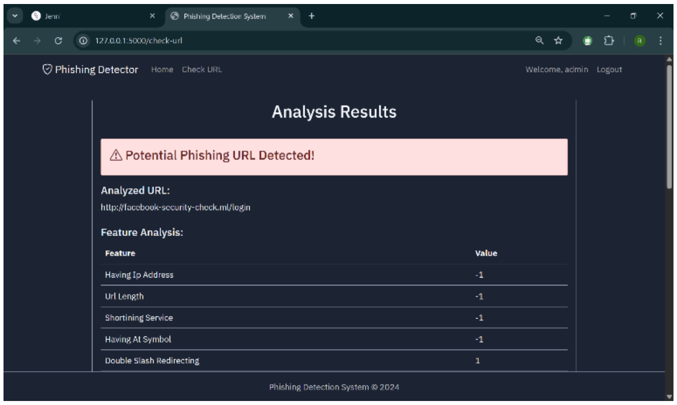
Feature	Value
Having Ip Address	-1
Url Length	-1
Shortining Service	-1
Having At Symbol	-1
Double Slash Redirecting	-1

Phishing Detection System © 2024


Rohith28h (rohith.h) · GitHub
Rohith28h (rohith.h) · GitHub
Phishing Detection System
127.0.0.1:5000/check-url

Prefix Suffix	-1
Having Sub Domain	0
Sslfinal State	1
Domain Registeration Length	1
Favicon	1
Port	-1
Https Token	-1
Request Url	-1
Url Of Anchor	-1
Links In Tags	-1
Sfh	-1
Submitting To Email	-1
Abnormal Url	-1
Redirect	-1
On Mouseover	-1

Phishing Detection System © 2024



Analysis Results

⚠ Potential Phishing URL Detected!

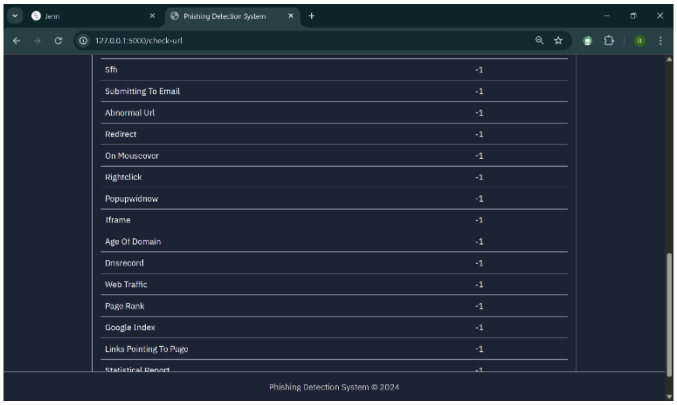
Analyzed URL:

http://facebook-security-check.ml/login

Feature Analysis:

Feature	Value
Having Ip Address	-1
Url Length	-1
Shortining Service	-1
Having At Symbol	-1
Double Slash Redirecting	1

Phishing Detection System © 2024



Sfh	-1
Submitting To Email	-1
Abnormal Url	-1
Redirect	-1
On Mouseover	-1
Rightclick	-1
Popupwidnow	-1
Iframe	-1
Age Of Domain	-1
Dnsrecord	-1
Web Traffic	-1
Page Rank	-1
Google Index	-1
Links Pointing To Page	-1
Statistical Report	-1

Phishing Detection System © 2024

V. Conclusion

The proposed machine learning framework provides an effective solution for detecting phishing websites. By leveraging advanced machine learning models and automated feature extraction, the system can accurately classify URLs as phishing or legitimate, reducing the risk of malicious web deception. The framework is designed to be adaptable to evolving phishing techniques, ensuring high accuracy and reliability in real-time detection. Future work will focus on enhancing the system's ability to detect sophisticated phishing attacks and improving its overall performance.

VI. Future Work

Future improvements will focus on enhancing the system's ability to detect highly sophisticated phishing techniques. This includes integrating more advanced machine learning models, such as deep learning, and improving the feature extraction process to capture more nuanced patterns in URLs. Additionally, the system will be optimized for real-time processing, allowing it to handle large volumes of data efficiently. Further research will also explore the integration of additional data sources, such as website content and user behavior, to improve the accuracy of phishing detection.

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. Sanchez-Gordon, 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 Learning-Based Approach," in 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), pp. 0250-0255, 2021