Secure URL: A Phishing Website Detection Application

Dr. T. Adilakshmi1, Dr. T. Jalaja2, Harish Reddy Ch 3, Gayathri B4

[t\_adilakshmi@staff.vce.ac.in](mailto:hodcse@staff.vce.ac.in)1, [jalaja.t@staff.vce.ac.in](mailto:jalaja.t@staff.vce.ac.in1)2 ,[chinthakuntlaharish@gmail.com](mailto:chinthakuntlaharish@gmail.com)3, [gayatribelide@gmail.com](mailto:gayatribelide@gmail.com)4

1,2,3,4 Department of Computer Science and Engineering, Vasavi College of Engineering, Hyderabad, India

***Abstract-*** Phishing is one of the most prevalent forms of social engineering and cyberattacks, targeting unsuspecting online users by tricking them into disclosing sensitive information for fraudulent purposes. These attacks pose a significant threat to online security and highlight the urgent need for robust detection mechanisms. Traditional approaches like user awareness campaigns and blacklisting phishing websites have limitations, as they often fail to identify phishing sites early or adapt to newly emerging threats.

This project leverages data mining techniques, specifically machine learning and deep neural network algorithms, to detect phishing websites at their early stages. By training models on a dataset comprising both phishing and benign website URLs, along with their content-based features, we aim to create an intelligent system capable of accurately identifying phishing sites.

The primary goal of this research is to evaluate the performance of various machine learning algorithms and neural networks in detecting phishing websites and to identify the most effective approach. Through feature extraction and rigorous performance evaluation, the project seeks to develop a reliable, scalable, and automated phishing detection system to mitigate the risks posed by such cyber threats.

**Keywords-** Phishing Detection, Machine Learning, SSL Analysis, URL Feature Extraction, Cybersecurity, Gradient Boosting, Zero-Day Attacks.

**I. INTRODUCTION**

Phishing is a sophisticated cyber threat that has become increasingly prevalent in the digital age, targeting online users by mimicking legitimate websites to steal sensitive information such as usernames, passwords, and financial details. These attacks exploit users' trust and are a common method of social engineering. Despite the availability of security awareness programs and traditional solutions like maintaining blacklists of phishing sites, these approaches often fail to prevent attacks on newly created phishing websites due to their dynamic and adaptive nature.

To address these challenges, advanced data-driven techniques such as machine learning (ML) and deep learning (DL) have emerged as effective solutions. These techniques allow for the analysis of vast datasets containing URL patterns, website content, and other indicators to identify phishing websites with high accuracy. By focusing on URL-based features, website metadata, and behavioral patterns, ML models and deep neural networks can detect phishing websites at an early stage, even before significant harm is done.

This project aims to develop a phishing website detection system by leveraging data mining methods, training machine learning models, and evaluating their performance. By analyzing a dataset comprising both phishing and benign websites, key features are extracted, and the system is trained to predict phishing attempts. This approach not only improves detection rates but also provides a scalable solution for mitigating phishing attacks.

**II.RELATED WORK**

[1] List-Based Detection Methods for Phishing Websites: This study explores list-based detection mechanisms, which include maintaining blacklists of known phishing URLs and whitelists of trusted URLs. Blacklists, used in browsers like Google Chrome and Mozilla Firefox via Google Safe Browsing, provide warnings when malicious sites are detected. Whitelists, on the other hand, help identify trusted websites by flagging URLs not present in the list as suspicious.

Limitation:List creation and updates should be based on lightweight mechanisms not to introduce delays in the detection process,Lists should be constantly updated to defend against newly discovered phishing attacks, Rules and heuristics devised for creating and updating the lists should reflect in a timely manner the evolution of the tactics adopted by attackers

[2]Page similarity based detection:The paper explores page similarity methods for detecting phishing websites. It compares suspicious pages to legitimate ones based on **textual** (HTML, CSS, DOM) and **visual** (images, logos) content. The similarity scores help identify phishing attempts by comparing these elements, with final decisions based on predefined thresholds.

**Limitations:** Effectiveness**(**Evasion techniques like code obfuscation and image distortions can reduce detection accuracy.), Approaches relying on external services can be slow.,Large datasets of legitimate pages require significant storage capacity

[3] The study outlines various phishing techniques, including email phishing, spear phishing, whaling, smishing, vishing, clone phishing, and social media phishing. It emphasizes the sophistication of modern attacks, which leverage social engineering and advanced technologies like AI.

**Limitations:** Although the paper provides a broad review, it lacks detailed explanations of the technical implementation of detection and prevention techniques, such as specific algorithms or systems.

**III. PROPOSED METHODOLOGY**

**Proposed System Overview**

The proposed system is designed to detect phishing websites using a combination of machine learning and deep learning models. It involves the following components and processes:

**Data Collection:**A dataset of phishing and legitimate URLs is collected from reliable sources.The dataset includes both static (URL-based) and dynamic (content-based) features to ensure comprehensive analysis.

**Feature Extraction:**URL-based features such as domain age, IP address usage, length of the URL, and presence of suspicious keywords.Content-based features, including HTML content, JavaScript behavior, and SSL certificate details.Behavioral indicators like website redirection patterns and user interaction logs.

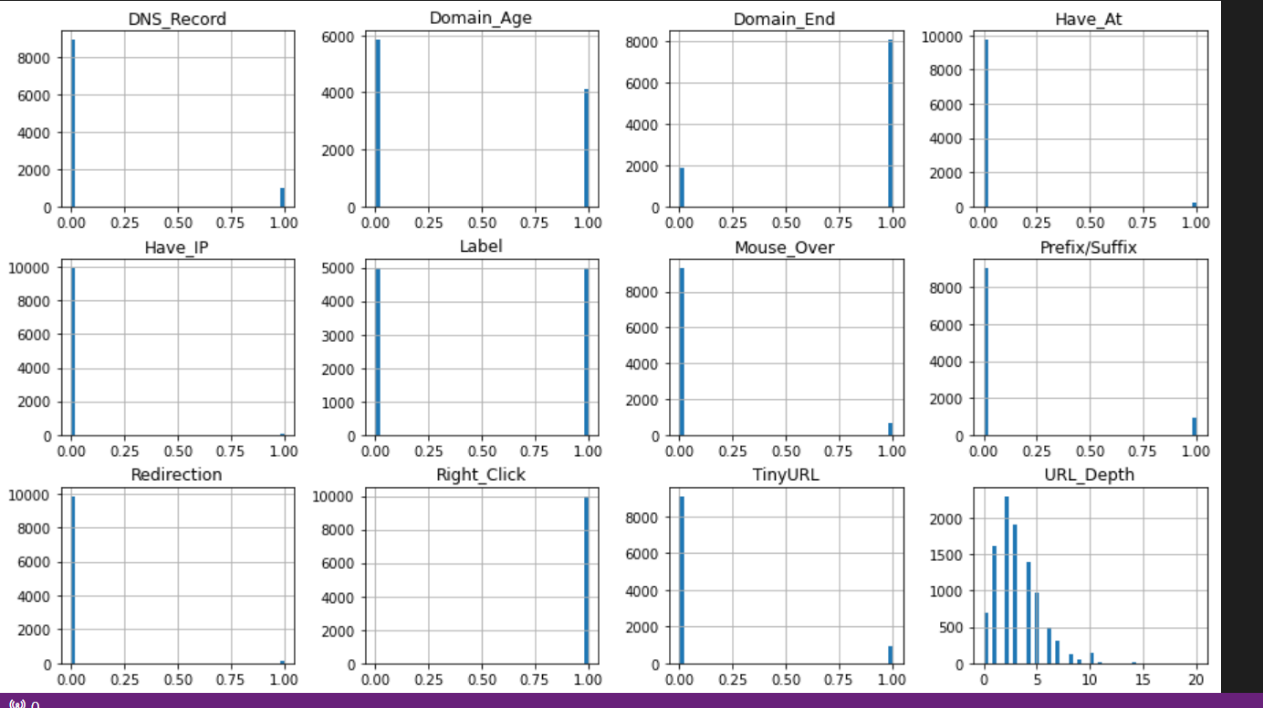
**Data Preprocessing:**The collected data is cleaned, normalized, and encoded for machine learning compatibility.Missing values are handled, and irrelevant features are removed to optimize model performance.

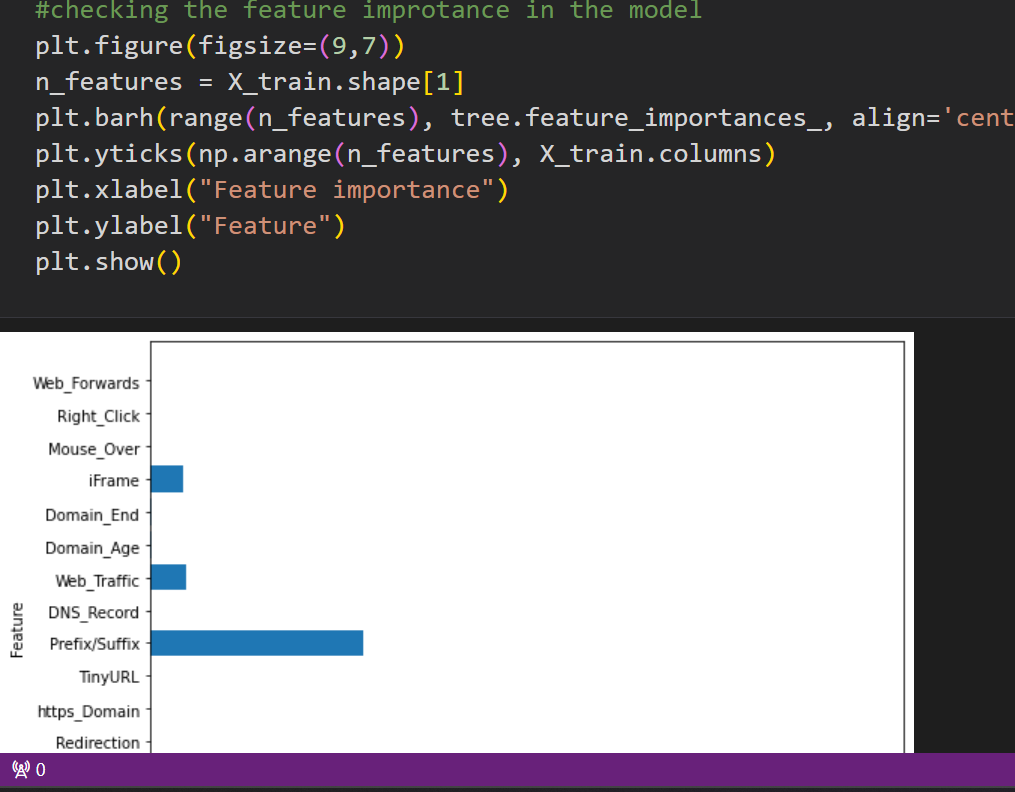
**Model Training and Evaluation:**Multiple machine learning models (e.g., Decision Trees, Random Forests, and Support Vector Machines) are trained using the dataset.Deep learning architectures (e.g., Neural Networks) are used for enhanced feature learning.Models are evaluated using metrics such as accuracy, precision, recall, and F1-score.A trained model predicts whether a given website is phishing or legitimate based on extracted features.The system is designed to adapt to new data through incremental learning, enhancing its ability to detect novel phishing techniques.

**Performance Comparison:**The performance of various models is compared to identify the most effective algorithm for phishing detection.

**User Interface:**A user-friendly interface is provided for end-users to input URLs and receive real-time predictions.

**IV. RESULTS AND ANALYSIS**

****

****

A screenshot of a graph

AI-generated content may be incorrect.

**V. CONCLUSION AND FUTURE SCOPE**

In conclusion, the phishing website detection project successfully demonstrates the application of machine learning algorithms to identify and classify websites as either phishing or legitimate. By leveraging features extracted from URLs, website content, and other indicators, the system provides a robust and efficient solution for mitigating phishing attacks.

Key achievements include:

1. **High Accuracy:** The model achieves a commendable accuracy, showcasing its reliability in real-world scenarios.
2. **Feature Effectiveness:** The selected features, such as URL characteristics, domain information,SSL and TLS certificate verifications and content-based features, contribute significantly to accurate predictions.

Future work could involve:

* Integrating the system with browsers or email filters for real-time phishing detection.
* Enhancing feature extraction techniques to improve model performance.
* Implementing advanced algorithms, such as deep learning or ensemble methods, for further accuracy improvements.
* Addressing multilingual and non-English phishing websites to expand applicability.

This project serves as a vital tool in the fight against cyber threats, offering a scalable and intelligent approach to enhance online security and protect users from phishing scams.

**VI. REFERENCES**

1)" Phishing or Not Phishing? A Survey on theDetection of Phishing Websites "

Rasha Zieni , Lusia Massari, and Maria Carla Calzarosa, (Senior Member,,IEEE),Department of Electrical, Computer and Biomedical Engineering, Università di Pavia, 27100 Pavia, Italy

2)A Review on Phishing Attacks “

Reference:Akarshita Shankar, Ramesh Shetty, International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 9 (2019) pp. 2171-2175

Source: IEEE Xplore.

3)https://blog.keras.io/building-autoencoders-in-keras.html

4)https://en.wikipedia.org/wiki/Autoencoder

5)https://mc.ai/a-beginners-guide-to-build-stacked-autoencoder-and-tying-weights-with-it/