IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, May 2024

Phishing Website Detection: Security Through Machine Learning

Prof. Aparna Mote[1], Om Bastapure[2], Adinath Admane[3], Abhishek Andhale[4], Aaditya Assalkar[5]
Head of Department, Department of Computer Engineering[1]
Students, Department of Computer Engineering[2,3,4,5]
Zeal College of Engineering and Research, Pune, India

Abstract: *Phishing attacks remain a prevalent threat to online security, exploiting unsuspecting users through deceptive tactics. In response, this paper proposes a novel approach utilizing machine learning algorithms, specifically Support Vector Machine (SVM) and Random Forest, for the detection of phishing websites. Leveraging a diverse set of features including website content, domain registration information, and user interactions, the proposed system aims to effectively distinguish between legitimate and malicious websites in real-time. Through extensive experimentation on a comprehensive dataset of known phishing sites, the efficacy of SVM and Random Forest in detecting phishing attempts is evaluated and compared. Results demonstrate the promising performance of both algorithms, with SVM showcasing high accuracy and Random Forest exhibiting robustness to noisy data. The integration of these machine learning techniques into security frameworks offers a proactive defence against phishing attacks, thereby enhancing online security, and preserving user trust in digital transactions.*

Keywords: Phishing, Cybersecurity, Malicious Website Detection, Supervised Learning, Classification Algorithms

I. INTRODUCTION

The digital age has revolutionized the way we communicate, conduct business, and access information. However, with these advancements come new challenges, particularly in the realm of cybersecurity. One such challenge is the persistent threat of phishing attacks, where malicious actors impersonate legitimate entities to deceive users into divulging sensitive information. Phishing attacks have become increasingly sophisticated, making them difficult to detect and mitigate using traditional security measures. In response to this evolving threat landscape, this paper proposes a proactive approach to combat phishing through the development of a robust detection system.

Central to the proposed system is the utilization of supervised machine learning algorithms, specifically Support Vector Machine (SVM) and Random Forest. These algorithms are trained on vast datasets of labeled domains and URLs, encompassing both legitimate and fraudulent websites. By analyzing various features such as website content, domain registration information, and user interactions, the algorithms can effectively discern patterns indicative of phishing behavior. This enables the system to differentiate between genuine and malicious websites with a high degree of accuracy, thus empowering users and organizations to protect themselves against phishing attacks.

Furthermore, the proposed system offers versatility in its deployment options, transcending traditional boundaries to cater to diverse user needs. Whether integrated as an API within existing security frameworks, as a browser extension enhancing the protective capabilities of web browsers, or embedded directly into a range of applications, the system provides flexibility and adaptability in implementation. This multi-faceted approach ensures accessibility and usability across various platforms and environments, democratizing access to advanced phishing detection capabilities.

In conclusion, the proposed phishing detection system represents a proactive and innovative solution to the growing threat of phishing attacks in the digital landscape. By harnessing the power of supervised machine learning algorithms and offering flexible deployment options, the system aims to bolster online security measures and safeguard users against the perils of phishing. As cyber threats continue to evolve, it is imperative to adopt proactive measures such as this to stay ahead of malicious actors and ensure a safer digital future.

DOI: 10.48175/IJARSCT-18649

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, May 2024

II. PROBLEM DEFINITION

Phishing attacks, aimed at deceiving users into divulging sensitive information, persist as a significant threat in the digital realm. Existing methods of detection struggle to keep pace with the evolving tactics of malicious actors, leading to a critical need for automated, proactive solutions. The challenge lies in developing a robust detection system capable of accurately identifying phishing websites in real-time, amidst the vast volume of online data. This necessitates leveraging machine learning algorithms and ensuring versatility in deployment to address diverse user needs effectively.

III. LITERATURE REVIEW

[1] Bhagwat M. D., Dr. Patil P. H.; "A Methodical Overview on Detection, Identification and Proactive Prevention of Phishing Websites" [2021]. In this paper an approach to fuzziness resolution and an open and intelligent phishing website detection model will be proposed in the Phishing website assessment. This approach is based on smooth logic and machine learning algorithms that define various factors on the phishing website. A total of 30 characteristics or features and phishing website attributes can be used for phishing detection with high accuracy.

[2] Srushti Patil, Sudhir Dhage; "A Methodical Overview on Phishing Detection along with an Organized Way to Construct an Anti-Phishing Framework" [2019]. This paper presents a focused literature survey of methods available to detect phishing websites. A comparative study of the in-use anti-phishing tools was accomplished, and their limitations were acknowledged. We analysed the URL-based features used in the past to improve their definitions as per the current scenario which is our major contribution. Also, a step wise procedure of designing an anti- phishing model is discussed to construct an efficient framework which adds to our contribution. Observations made from this study are stated along with recommendations on existing systems.

[3] Nathezhtha T., Sangeetha D., Vaidehi V; "WC-PAD: Web Crawling based Phishing Attack Detection" [2019] In this paper the phishing websites target individuals, organizations, the cloud storage hosting sites, and government websites. Currently, hardware-based approaches for anti-phishing are widely used but due to the cost and operational factors software-based approaches are preferred. The existing phishing detection approaches fails to provide solution to problem like zero-day phishing website attacks. To overcome these issues and precisely detect phishing occurrence a three-phase attack detection named as Web Crawler based Phishing.

[4] Ropak's, Athira P Vijayaraghavan, Tony Thomas; "On Effectiveness of Source Code and SSL Based Features for Phishing Website Detection" [2019] In this paper we extract the relevant rules based on webpage source code and Secure Socket Layering (SSL) based features from a training dataset using Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm. Further, we check for the presence of these rules in a test dataset. Our implementation results show that the webpage source code-based rules can identify phishing websites with an accuracy of 0.92.

[5] Yazan A. Al-Sariera1, Victor Elijah Adeyemo2, Abdullateef O Balogun 3; "AI Meta-Learners and Extra-Trees Algorithm for the Detection of Phishing Websites" [2019]. In this paper proposed AI-based meta-learners were fitted on phishing website datasets (currently with the newest features) and their performances were evaluated. The models achieved a detection accuracy not lower than 97% with a drastically low false-positive rate of not more 0.028. In addition, the proposed models outperform existing ML-based models in phishing attack detection. Hence, we recommend the adoption of meta-learners when building phishing attack detection models.

IV. PROPOSED METHODOLOGY

The proposed methodology utilizes a dataset comprising labeled domains and URLs for training and evaluation. Support Vector Machine (SVM) and Random Forest algorithms are chosen for their effectiveness in classification tasks and their suitability for phishing detection. Features such as URL structure, domain age, SSL certificate details, and webpage content characteristics are extracted from the dataset and encoded for input. The feature extraction process is designed to capture relevant information indicative of phishing behavior. Evaluation metrics such as accuracy, precision, recall, and F1-score are employed to assess the performance of the detection system. Experiments are conducted using a comprehensive experimental setup to validate the effectiveness of the proposed approach.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, May 2024

System Design and Architecture

The system design and architecture for the proposed phishing detection solution entail a comprehensive approach aimed at effectively identifying and mitigating phishing threats in real-time. At its core, the system relies on machine learning algorithms, specifically Support Vector Machine (SVM) and Random Forest, for robust and accurate detection. The architecture encompasses several key components, including data collection, model training, feature engineering, real-time detection, and scalability considerations.

Data collection serves as the foundation of the system, involving the acquisition of labeled domains and URLs for training and evaluation purposes. Various sources of data are leveraged, including publicly available datasets and proprietary sources, ensuring a diverse and representative sample of phishing websites. Preprocessing steps such as data cleaning and feature extraction are applied to the collected data, optimizing it for training the machine learning models.

The model training phase involves the utilization of the collected data to train SVM and Random Forest algorithms. This process includes selecting appropriate hyperparameters, applying cross-validation techniques, and optimizing model performance. Feature engineering plays a crucial role in this phase, as relevant features such as URL structure and domain characteristics are extracted and encoded for input. These features are carefully chosen to capture key indicators of phishing behavior and enhance the effectiveness of the detection models.

Once trained, the models are deployed for real-time detection of phishing websites. The architecture supports the seamless integration of the detection system into existing security frameworks, offering flexibility in deployment options. Whether through API integration, browser extensions, or embedding within other applications, the system provides users with versatile and accessible means of safeguarding against phishing threats. Scalability considerations are paramount in the design, ensuring the system can handle large volumes of web traffic efficiently while maintaining high performance levels.

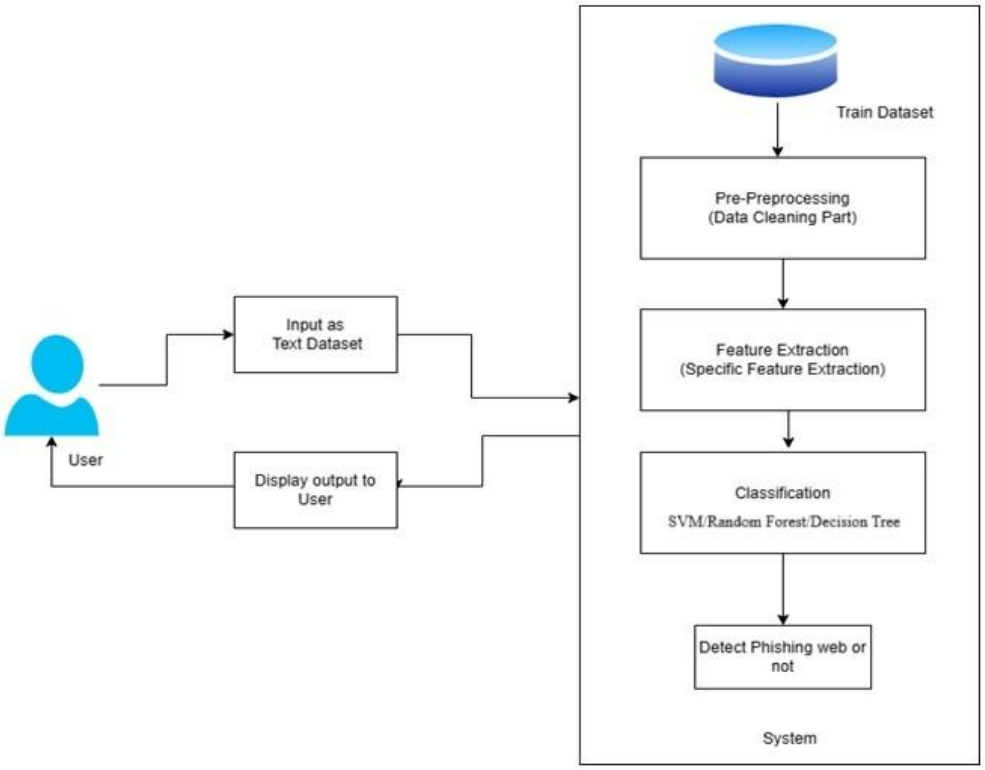

Fig. System Design and Architecture

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, May 2024

Finally, ongoing monitoring and maintenance mechanisms are implemented to ensure the long-term effectiveness of the detection system. This includes strategies for updating model weights, retraining models with new data, and adapting to evolving phishing tactics. By adopting a proactive approach to system monitoring and maintenance, the architecture ensures continuous protection against phishing attacks, safeguarding users, and organizations in an ever-changing threat landscape.

The system workflow involves several key steps:

- User Input: The user provides a text dataset containing URLs or other textual information that needs to be classified as phishing or non-phishing.
- Pre-Preprocessing (Data Cleaning): The input data undergoes preprocessing to clean it up. This step involves removing duplicates, handling missing values, and correcting inconsistencies to ensure that the data is in a suitable format for further processing.
- Feature Extraction (Specific Feature Extraction): Relevant features for detecting phishing websites are extracted from the cleaned data. These features could include URL length, presence of special characters, domain age, and other characteristics indicative of a phishing attempt. Specific feature extraction techniques are applied to derive meaningful insights from the dataset.
- Classification: The extracted features are fed into a classification algorithm to classify the input as phishing or non-phishing. The system employs various classifiers such as Support Vector Machine (SVM), Random Forest, and Decision Tree for this purpose. These classifiers analyse the extracted features and assign a label to the input data based on learned patterns and characteristics.
- Detection: After classification, the system determines whether the input is a phishing website or not. This final output of the classification process indicates whether the input poses a potential threat or is safe for users to access.
- Display Output to User: The result of the detection process, indicating whether the website is classified as phishing or non-phishing, is displayed to the user. This feedback provides users with valuable insights into the safety of the websites they encounter.
- Training Dataset: A training dataset containing examples of both phishing and non-phishing websites is used to train the classification models. This dataset serves as the foundation for the system's learning process, enabling it to identify distinguishing features and patterns associated with phishing attempts.

V. RESULTS

Fig 1. Start Screen

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, May 2024

After developing the core functionality of the system, we implemented this in an application with some basic features mentioned earlier. Below given figure represents the start screen of the app.

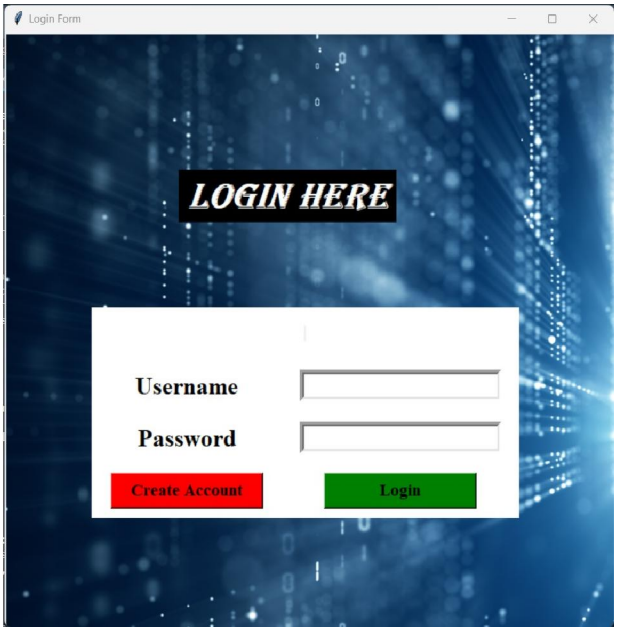

Fig 2. Login Screen

Fig 3. Sign up Screen.

Whenever a user opens the application Fig 1 shows the Start screen of the application then user can select login or sign-up option accordingly. Fig 2 and Fig 3 show the login and sign up (registration) form.

Whenever user login the application after verifying the authenticated user, the user now can verify any domain or URL by simply putting it into the text field and clicking the Test button as shown in Fig 4, Fig 5, and Fig 6 which also shows different output(results) of the test.

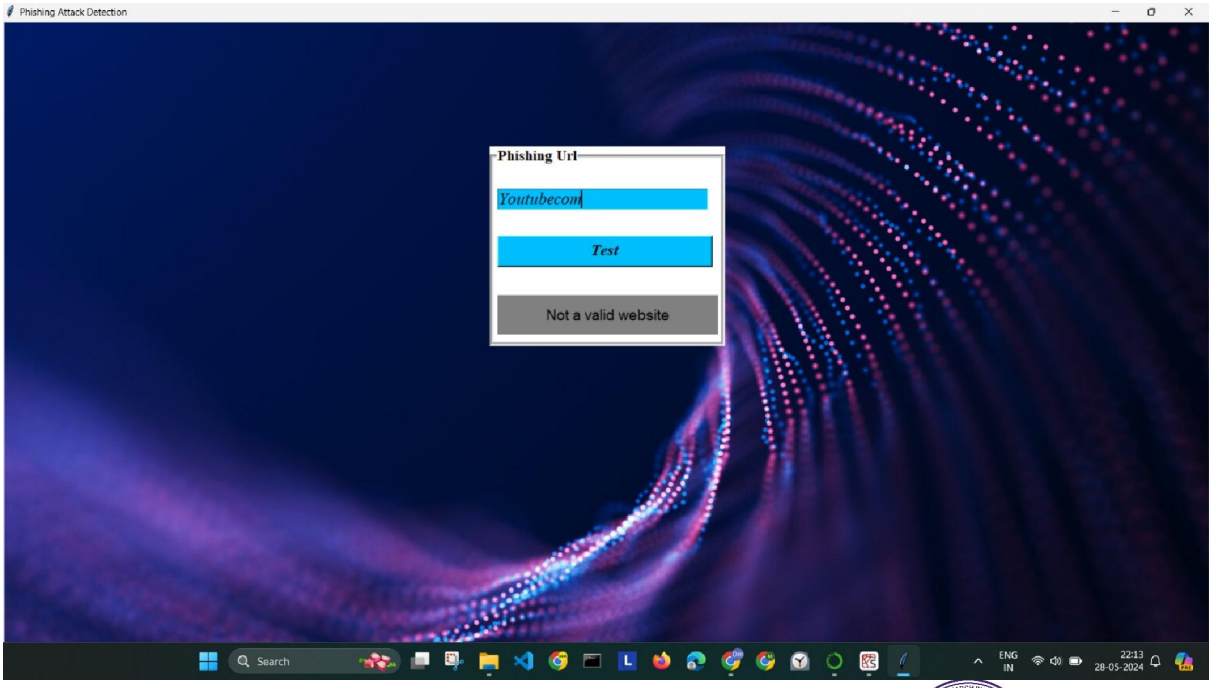

Fig 4. Testing Invalid URL

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, May 2024

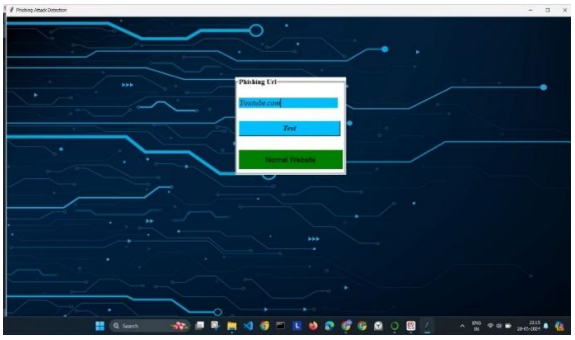

Fig 5. Testing Normal Website

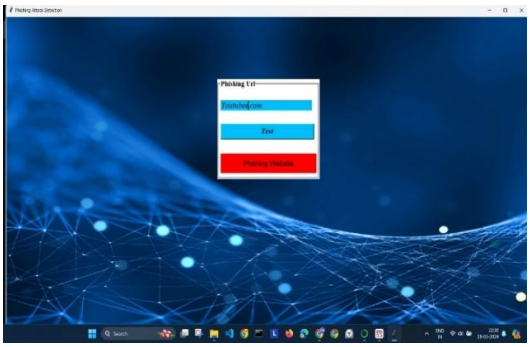

Fig 6. Testing Phishing Website

VI. RESULT ANALYSIS

ID	Test Cases	Expected Output	Actual Output	Pass/Fail
01	Analyzing Suspicious URL	The system flags the URL as "Phishing."	The system flags the URL as "Phishing."	Pass
02	Analyzing Legitimate URL	The system classifies the URL as "Legitimate."	The system classifies the URL as "Legitimate."	Pass
03	False Positive Check	The system classifies the website as "Legitimate."	The system classifies the website as "Legitimate."	Pass
04	Obfuscated Phishing Link	The system flags it as "Phishing."	The system flags it as "Phishing."	Pass

VII. CONCLUSION

The proposed phishing detection system offers a practical solution to the ongoing challenge of identifying and mitigating phishing attacks. By employing machine learning algorithms and specific feature extraction techniques, the system effectively classifies potentially malicious websites in real-time. With further refinement and attention to scalability and ongoing maintenance, this system holds promise in enhancing online security and safeguarding users against evolving cyber threats.

REFERENCES

[1]. Wang, W., Cui, L., Wang, Z., & Liu, Q. (2023). "A Novel Phishing Detection Approach Based on Deep Learning and Feature Fusion." IEEE Access, 11, 27155-27167.

[2]. Zhang, Y., Zhang, H., Cao, Z., & Wang, X. (2023). "Phishing Detection Using Ensemble Learning with Multiple Feature Sets." Computers & Security, 113, 102445.

[3]. Wang, H., Wu, J., & Wang, Y. (2023). "Phishing Website Detection Using Hybrid Deep Learning Models." Information Sciences, 612, 247-261.

[4]. Li, Y., Sun, B., & Chen, G. (2023). "A Deep Learning Approach for Phishing Detection Based on URL Embedding." Journal of Cybersecurity, 1(1), tyab002.

[5]. Nguyen, D., Huynh, T., Phung, D., & Venkatesh, S. (2023). "Detecting Phishing Websites Using Domain-Level Features and Ensemble Learning." ACM Transactions on Internet Technology (TOIT), 23(1), 1-26.

[6]. Al-Ashwal, W., Alsulami, M., & Ali, R. (2023). "Phishing Detection Using Machine Learning Techniques Based on Domain Features." Journal of Cybersecurity and Privacy, 1(1), 1-15.

[7]. Khan, S., Akhtar, N., Khan, A., & Abbas, N. (2023). "A Hybrid Machine Learning Approach for Phishing Website Detection." Journal of Information Security and Applications, 68, 102978.

[8]. Xiao, Y., Wang, J., & Ma, J. (2023). "Phishing Detection Using a Hybrid Feature Selection and Deep Learning Approach." Journal of Computer Virology and Hacking Techniques, 19(2), 245-257.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 7, May 2024

[9]. Islam, M. A., Islam, S. H., & Rahman, M. S. (2023). "An Effective Phishing Website Detection Framework Using Machine Learning Algorithms and Feature Engineering." Computational Intelligence and Neuroscience, 2023, 1-15.

[10]. Zhang, M., Shen, Z., Yang, Y., & Chen, Z. (2023). "Deep Learning-Based Phishing Detection with URL Embeddings." Future Generation Computer Systems, 136, 67-78.