Phishing Detection System Using Machine Learning

Mr. Anirudh Shitole, UG Student SKN Sinhgad Institute of Technology & Science, Lonavala, Maharashtra, India, anirudhshitole.sknsits.comp@gmail.com

Mr. Akash Shitole, UG Student SKN Sinhgad Institute of Technology & Science, Lonavala, Maharashtra, India, akashshitole.sknsits.comp@gmail.com

Mr. Aniket Shelke, UG Student SKN Sinhgad Institute of Technology & Science, Lonavala, Maharashtra, India, aniketshelke.sknsits.comp@gmail.com

Mr. Manas Joshi, UG Student SKN Sinhgad Institute of Technology & Science, Lonavala, Maharashtra, India, manasjoshi.sknsits.comp@gmail.com

Mr. S. P. Gunjal, Asst. Professor SKN Sinhgad Institute of Technology & Science, Lonavala, Maharashtra, India, spgunjal.sknsits@sinhgad.edu

Abstract: Phishing remains a significant cybersecurity threat, where attackers mimic legitimate websites to deceive users into divulging sensitive information. Detecting phishing websites accurately and promptly is paramount to mitigate potential risks. In this study, we propose a robust phishing website detection system leveraging machine learning techniques. Our system integrates a diverse set of features including URL-based features, website content analysis. The system contains machine learning approaches such as gradient boosting classifier, logistic regression which trains the system to predict whether the website is safe for use or not.

Keywords — Cybersecurity, Gradient Boosting Classifier, Logistic Regression, Machine learning, Phishing

I. INTRODUCTION

In today's digital landscape, where online activities have become an integral part of our daily lives, the threat of phishing attacks looms large. Phishing attacks involve malicious entities impersonating legitimate websites or entities to deceive unsuspecting users into divulging sensitive information such as login credentials, financial details, or personal data. These attacks not only pose significant risks to individuals but also threaten the security and integrity of organizations and their data.

Traditional methods of detecting phishing websites often rely on manually curated blacklists, which are unable to keep pace with the rapidly evolving nature of phishing techniques. As a result, there is a pressing need for more sophisticated and automated approaches to detect phishing websites effectively.

Machine learning, a branch of artificial intelligence that enables systems to learn from data and improve over time without being explicitly programmed, presents a promising solution to this challenge. By leveraging machine learning algorithms, it is possible to develop robust systems capable of analyzing large datasets of features extracted from websites to distinguish between legitimate and phishing websites.

This system contributes to the ongoing efforts to bolster cybersecurity defenses against phishing threats, ultimately fostering a safer and more secure online environment for users and organizations.

II. BACKGROUND STUDY

Phishing Trends and Challenges: Begin by reviewing studies that highlight the evolving trends and challenges in phishing attacks. Understand the tactics used by attackers to create deceptive websites and lure users into divulging sensitive information. This background knowledge will provide context for the importance of effective detection systems.

Traditional Detection Methods: Review existing approaches to phishing detection, including heuristic-based methods, blacklists, and manual inspection of suspicious websites. Understand the limitations of these methods, such as their reliance on static signatures and their inability to adapt to evolving phishing tactics.

Machine Learning in Phishing Detection: Explore the growing body of research on applying machine learning techniques to detect phishing websites. Investigate studies that leverage features extracted from website content, structure, metadata, and user behavior to train classifiers capable of distinguishing between legitimate and phishing sites.

Feature Engineering: Delve into feature engineering

 DOI : 10.35291/2454-9150.2024.0186

methodologies tailored for phishing detection. Examine the types of features commonly used in phishing detection systems, such as URL-based features (e.g., domain age, presence of subdomains), content-based features (e.g., lexical analysis, HTML attributes), and behavioral features (e.g., mouse movements, click patterns).

Challenges and Future Directions: Consider the challenges and open research questions in phishing detection, such as detecting zero-day phishing attacks, handling imbalanced datasets, and scaling detection systems to handle large volumes of web traffic. Identify emerging trends and promising directions for future research in the field.

III. LITERATURE SURVEY

In this paper, we will discuss previous studies in terms of their methods, datasets, contributions, and results.

S. Arvind Anwekar, V. Agrawal [19]: In this study, the authors focused on extracting features from URLs, in addition to other features such as the age of the SSL certificate and the universal resource locator of the anchor, IFRAME, and website rank. They collected URLs of phishing websites from PhishTank and URLs of benign websites from the Alexa website. Using a combination of the random forest (RF), decision tree and support vector machine (SVM).

N. Choudhary b, K. Jain, S. Jain [20]: This study emphasizes the significance of only using attributes from the URL. Both the Kaggle and Phishtank websites make it easy to get the dataset used in this study. The researchers used a hybrid approach that combined Principal Component Analysis (PCA) with Support Vector Machine (SVM) and Random Forest algorithms.

A. Lakshmanarao, P. Surya, M Bala Krishna [21]: This thesis collected a dataset of phishing websites from the UCI repository and used various Machine learning techniques, including decision trees, AdaBoost, support vector machines (SVM), and random forests, to analyze selected features (such as web traffic, port, URL length, IP address.

IV. METHODOLOGY

Detecting phishing websites involves a combination of techniques and methodologies to accurately identify malicious sites and protect users from potential threats. Here's a suggested methodology for phishing website detection using a machine learning approach:

1. Data Collection:

Gather a diverse dataset of websites, including both legitimate and known phishing sites. Ensure the dataset represents various industries and types of websites to create a robust model. Publicly available datasets like the Phishing Websites dataset (such as the one from UCI Machine Learning Repository) can be used, along with additional data scraped from the web.

2. Feature Extraction:

Extract relevant features from the URLs, website content, and meta-information. Features may include:

URL Features: Length of the URL, presence of '@' or '-', use of IP address, presence of 'https,' etc.

Content Features: Keywords, phrases, or patterns often found in phishing websites. Use techniques like TF-IDF (Term Frequency-Inverse Document Frequency) to extract relevant content features.

Meta-information Features: Extract metadata such as domain age, SSL certificate details, and registrar information.

3. Data Preprocessing:

Cleanse the dataset by handling missing values, normalizing feature values, and encoding categorical variables. Data preprocessing ensures that the dataset is suitable for machine learning algorithms.

4. Model Selection:

Experiment with various machine learning algorithms suitable for binary classification tasks. Common algorithms for phishing website detection include Decision Trees, Random Forest, Support Vector Machines, and Neural Networks. Evaluate the performance of each algorithm using cross-validation techniques and choose the one that provides the best results.

5. Feature Selection:

Utilize feature selection methods like Recursive Feature Elimination (RFE) or feature importance scores from ensemble methods to identify the most relevant features. Removing irrelevant or redundant features can improve the model's accuracy and reduce computation time.

6. Model Training and Validation:

Split the dataset into training and validation sets. Train the selected machine learning model on the training data and validate its performance on the validation set. Use appropriate metrics such as accuracy, precision, recall, F1-score, and ROC AUC to evaluate the model's effectiveness in phishing website detection.

7. Testing and Deployment:

Evaluate the final model on a separate test dataset to assess its real-world performance accurately. Once satisfied with the results, deploy the model into the production environment where it can be used to detect phishing websites in real-time.

8. Continuous Monitoring and Updating:

Phishing techniques evolve over time, so it's crucial to continuously monitor the model's performance. Regularly update the dataset and retrain the model with new data to ensure it remains effective against emerging phishing threats.

 DOI : 10.35291/2454-9150.2024.0186

V. ARCHITECTURAL DIAGRAM

In this project workflow, first of all the user will login into the system with his credentials after that user will get the webpage which will ask to upload a dataset.

After uploading dataset the machine learning model will process this data and extract the useful features from it.

After that user is able upload URL of the website and model will analyze the components present in URL and provide the output that this website is safe or not for use.

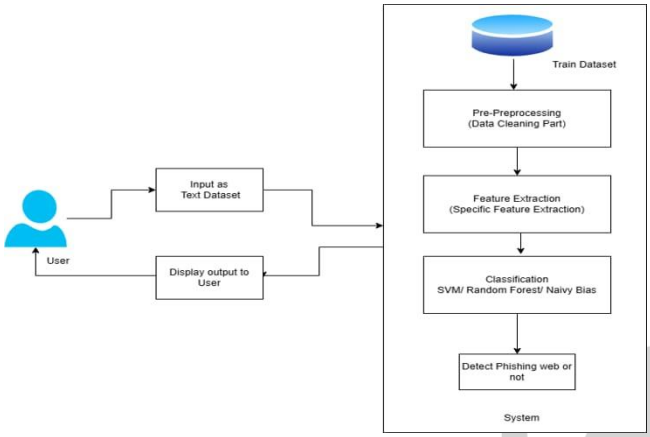

Fig V.1

VI. RESULTS

During this project we tried to create accurate, fast and effective application which can detect phishing websites, along with accuracy of model we also tried to build easy to use and responsive interface.

Phishing detection system will help individual in following possible ways:

Reduced Phishing Attacks:

The primary goal of the system is to identify and block phishing websites, leading to a decrease in successful phishing attacks targeting users and organizations.

Improved Security Posture:

With fewer successful phishing attempts, the overall security posture of the organization improves. This can result in fewer incidents of data breaches, financial losses, and reputational damage.

Protection of Sensitive Information:

By preventing users from accessing phishing websites, the system helps protect sensitive information such as login credentials, financial data, and personal details from falling into the hands of malicious actors.

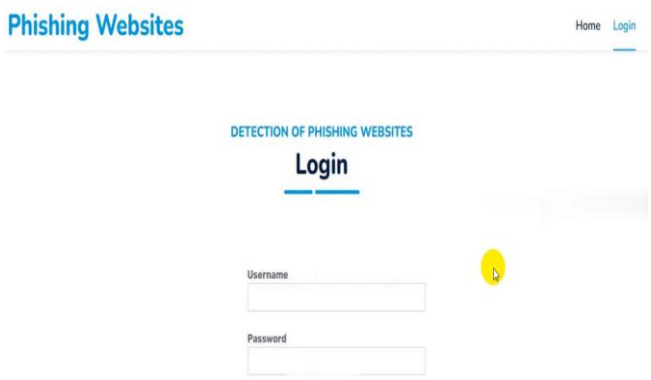

Fig VI.2

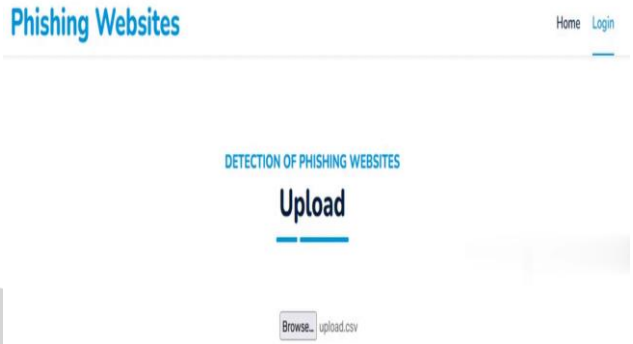

Fig VI.3

Fig VI.4

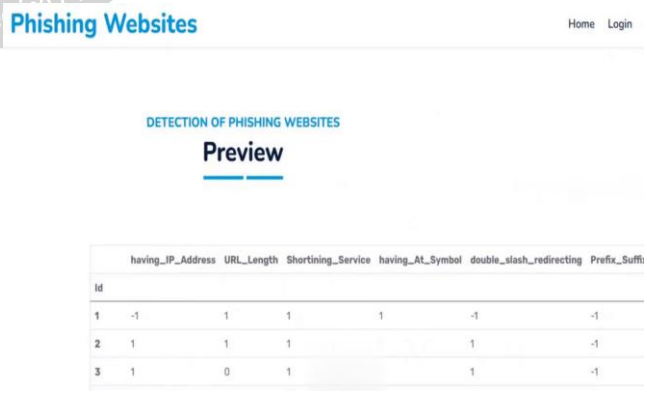

Fig VI.5

 DOI : 10.35291/2454-9150.2024.0186

Fig VI.6

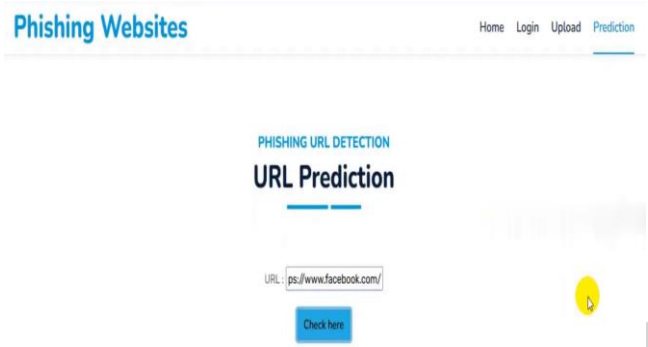

Fig VI.7

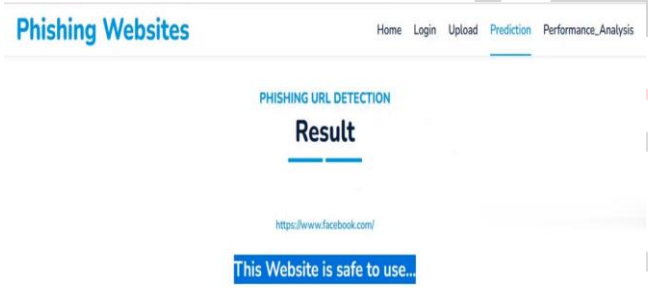

Fig VI.8

VII. CONCLUSION

Phishing website detection remains a critical area of research and development in the field of cybersecurity. With the increasing sophistication of phishing attacks and the ever-expanding threat landscape, effective detection systems are essential to protect individuals and organizations from falling victim to malicious schemes.

phishing detection systems play a crucial role in safeguarding individuals and organizations against one of the most prevalent cybersecurity threats: phishing attacks. These systems employ a combination of technologies, techniques, and human intelligence to identify and mitigate phishing attempts effectively.

Ultimately, the goal of phishing detection systems is to empower individuals and organizations to detect and avoid phishing attacks, thereby protecting sensitive information, financial assets, and reputation. While no system is foolproof, investing in robust phishing detection mechanisms is essential for maintaining a strong defense against this persistent and ever-evolving threat landscape.

In conclusion, by using this phishing detection system we are able to detect the websites that are unsafe for users and help them maintain their privacy.

VIII. REFERENCES

[1] A.Y. Ahmad, M. Selvakumar, A. Mohammed, and A.-S. Samer, “TrustQR: A new technique for the detection of phishing attacks on QR code,” Adv. Sci. Lett., vol. 22, no. 10, pp. 2905-2909, Oct.2021.

[2] C. C. Inez and F. Baruch, “Setting priorities in behavioral interventions: An application to reducing phishing risk,” Risk Anal., vol. 38, no. 4, pp. 826-838, Apr. 2021.

[3] Aburrous, Maher Hossain, Mohammed Dahal, Keshav Thabtah, Fadi. (2020). Intelligent phishing detection system for ebanking using fuzzy data mining. Expert Systems with Applications. 37. 7913-7921. 10.1016/j.eswa.2020.04.044.

[4]Rosiello, Angelo Kirda, Engin Kruegel, Ferrandi, Fabrizio. (2007). A layoutsimilarity-based approach for detecting phishing pages. Proceedings of the 3rd International Conference on Security and Privacy in Communication Networks, Secure Comm. 454 - 463. 10.1109/SECCOM..4550367.2021.

[5] Chawathe, Sudarshan. Improving Email Security with Fuzzy Rules. 1864-1869. 10.1109/TrustCom/BigDataSE.2018.00282. 2021.

[6] A. Aggarwal, A. Rajadesingan and P. Kumaraguru, ”PhishAri: Automatic realtime phishing detection on twitter,” eCrime Researchers Summit, Las Croabas, 2012, pp. 1-12, doi: 10.1109/eCrime.6489521 2022.

[7] P. Singh, Y. P. S. Maravi and S. Sharma, “Phishing Websites Detection through Supervised Learning Networks”, 2020 International Conference on Computing and Communications Technologies (ICCCT), Chennai, 2020, pp. 61-65.

[8] K. Thomas, C. Grier, J. Ma, V. Paxson and D. Song, “Design and Evaluation of a Real-Time URL Spam Filtering Service”, IEEE Symposium on Security and Privacy, Berkeley, CA, , pp. 447-462. 2021.

 DOI : 10.35291/2454-9150.2024.0186