IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 4, April 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304129 |

Detection of Phishing Websites

Dr. K. Shailaja[1], K. Vaishnavi[2], P. Shilpa[3], S. Naveen[4], CH. Uday Goud[5]

Assistant Professor, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India[1]

Student, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India[2]

Student, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India[3]

Student, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India[4]

Student, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India [5]

Abstract: In today's online world, phishing attacks, where scammers trick people into sharing personal info, are a big worry for everyone. To tackle this, we've created a smart system using machine learning. Instead of just reacting to attacks, our system stops them before they happen. We teach it what phishing looks like, so it can spot sneaky websites trying to trick you. By using fancy computer techniques like decision trees and neural networks, our system learns to tell real websites from fake ones. We've tested it thoroughly to make sure it works well in real situations and can handle different challenges. Our aim is to make the internet safer by staying one step ahead of cybercriminals, making it harder for them to fool people online and keeping users safe.

KEYWORDS: Uniform Resource Locator(Url), Phishing Urls, Legitimate Urls

I. INTRODUCTION

In today's digital landscape, the rise of phishing attacks poses a significant threat to online security. Phishing, a deceptive tactic in which scammers impersonate trusted entities to steal sensitive information, presents a formidable challenge for individuals and organizations alike. To address this issue, we've developed an innovative solution leveraging machine learning technology. Instead of merely responding to phishing attacks after they occur, our system takes a proactive approach, identifying and preventing them before they can do harm. Through extensive training on various phishing examples, our system learns to recognize the subtle cues indicative of fraudulent websites, bolstering our defenses against cyber threats.

Our solution represents a departure from traditional cybersecurity methods, emphasizing proactive detection and prevention strategies. By employing advanced computer algorithms such as decision trees and neural networks, our system possesses the intelligence needed to differentiate between legitimate and malicious websites. Rigorous testing, including real-world scenarios, validates the effectiveness and resilience of our solution, ensuring its ability to address the diverse challenges posed by cybercriminals. Ultimately, our objective is to enhance online safety by anticipating and countering phishing attacks, thereby protecting users from online fraud.

II. LITERATURE SURVEY

Alswailem et al. [1] explore the application of machine learning techniques for detecting phishing websites in their paper titled “Detecting Phishing Websites Using Machine Learning.”The process begins with feature extraction, where a thorough analysis of website attributes such as URL elements, HTML source code, and JavaScript code is conducted. This step aims to identify key indicators that distinguish phishing websites from legitimate ones. Subsequently, machine learning classification techniques, including algorithms such as Naive Bayes, Support Vector Machines (SVM), and Random Forest, are employed. These algorithms are trained on a labeled dataset comprising both phishing and legitimate websites, enabling them to learn and recognize patterns in website characteristics indicative of phishing behavior. Once the algorithms are trained, they are utilized for prediction, wherein new websites are classified based on the extracted features using the learned models. This predictive capability allows for the proactive identification and classification of potential phishing websites, thereby enhancing cybersecurity measures and mitigating the risks posed by online threats.The research by Razaque et al. [2] delve into the detection of phishing websites using machine

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304129 |

learning techniques in their study titled “Detection of Phishing Websites using Machine Learning,” presented at the 2020 IEEE Cloud Summit.The authors focus on implementing a comprehensive strategy to enhance cybersecurity measures, beginning with the establishment of blacklisting, which involves maintaining a database of known phishing websites to provide initial protection. Concurrently, they employ semantic analysis to scrutinize website content, including text, links, and images, for anomalies and suspicious keywords, thereby enhancing detection capabilities. Furthermore, the authors prioritize pattern recognition techniques, which entail the extraction of features such as URL length, presence of suspicious words, and domain extensions. These features are then subjected to machine learning algorithms to identify patterns, facilitating the detection of potential phishing attempts. To ensure real-time detection and proactive defense, the authors seamlessly integrate all these techniques into a Google Chrome extension. This extension empowers users with immediate protection as they browse the web, leveraging the combined capabilities of blacklisting, semantic analysis, and pattern recognition to safeguard against the evolving landscape of online threats.The literature on “Detecting Phishing Websites Using Machine Learning.” by M. H. Alkawaz and his collaborators[3] encompasses several key steps aimed at enhancing phishing detection capabilities. Initially, feature extraction involves a meticulous analysis of various website attributes, encompassing elements such as URL length, the presence of suspicious words, iframe tags, and visual elements like fonts and colors. Subsequently, feature selection employs statistical techniques to discern the most pertinent features for effective phishing detection, streamlining the analysis process. The subsequent phase involves machine learning classification, wherein algorithms such as Naive Bayes, Logistic Regression, and Support Vector Machines (SVM) are trained on a labeled dataset comprising both phishing and legitimate websites, utilizing the selected features. This training process enables the algorithms to discern patterns and characteristics indicative of phishing behavior. Finally, in the classification and prediction stage, newly encountered websites are evaluated based on their extracted features using the trained machine learning models, facilitating swift and accurate identification of potential phishing threats. This comprehensive approach leverages advanced statistical techniques and machine learning algorithms to bolster cybersecurity measures, ultimately enhancing protection against the pervasive threat of phishing attacks.

III. METHODOLOGY

The Feature Extraction class is designed to facilitate phishing detection by extracting relevant features from a given URL. Initially, the Using Ip method examines whether the URL is specified using an IP address directly or a domain name, returning a feature value to indicate its type. Subsequently, the long URL method evaluates the length of the URL, categorizing URLs based on their length to distinguish between suspicious and benign URLs. Additionally, the short URL method identifies if the URL is a shortened link by matching it against known URL shortening service patterns. The presence of specific symbols, such as the "@" symbol, in the URL is checked by the Symbol method, while the Redirecting method investigates if the URL contains multiple forward slashes ("//"), indicating possible redirection. The prefix Suffix method checks for hyphens ("-") in the domain, which could be indicative of suspicious domain names.

Further, the Subdomains method counts the number of subdomains in a URL to gauge its complexity and legitimacy. The Hppts method verifies if the URL uses the "https://" scheme, and the DomainRegLen method calculates the domain registration length based on WHOIS information. The presence of a favicon that matches the URL or domain is assessed by the Favicon method. Additionally, the NonStdPort method checks for the inclusion of a non-standard port in the URL, and the HTTPSDomainURL method examines the domain for the inclusion of "https", suggesting possible mimicking of the HTTPS protocol.

The Request URL method iterates through HTML elements, evaluating their URLs for specific patterns to categorize them as potentially malicious, needing further analysis, or likely legitimate. The AnchorURL method computes the percentage of unsafe anchor URLs, and the LinksInScriptTags method calculates the percentage of successful matches for links in 'link' and 'script' tags. The ServerFormHandler method analyzes HTML forms to determine if the 'action' attribute points to the current domain or an external location. Moreover, the InfoEmai method identifies the presence of email-related actions in the HTML content.

The Abnormal URL method compares the URL response text with the WHOIS response to determine its normality, and the Website Forwarding method analyzes the response history to assess the number of redirections encountered. The Status Bar Cust method searches for scripts handling the on mouseover event, potentially manipulating the status bar. The Disable Right Click and Using Pop up Window methods identify JavaScript scripts that disable the right-click

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304129 |

and create popup windows, respectively. The Iframe Redirection method detects iframe elements used for redirection. Additionally, the AgeofDomain method determines the domain's age based on its creation date from WHOIS data, and the DNS Recording method checks the DNS record configuration. The Website Traffic and PageRank methods assess the website's popularity using Alexa rank and page rank, respectively. Finally, the Google Index method searches for the website URL on Google, and the Links Pointing to Page method counts hyperlinks pointing to the webpage. The Stats Report method examines statistical reports to identify potential indicators of suspicious behavior, classifying the URLs based on the observed patterns.

The methodology for detecting phishing websites commences with the collection of a dataset from Kaggle[4], which contains various website attributes such as IP usage, URL characteristics, and security features. These attributes, including Index, UsingIP, LongURL, and others, serve as the basis for training machine learning models. Initially, several popular algorithms, namely XGBoost Classifier, Random Forest, Decision Tree, Logistic Regression, and Naive Bayes Classifier, are trained on the dataset to identify patterns indicative of phishing behavior. Following training, the XGBoost Classifier, exhibiting the highest accuracy among the models, is selected for further utilization. This model is serialized into a pickle file, facilitating easy integration into a front-end application. The front-end interface allows users to input a URL, prompting the system to determine whether the provided URL is a phishing or legitimate website. Behind the scenes, the system employs logic for feature extraction, analyzing the URL to extract relevant features such as URL length, presence of symbols, and security protocols like HTTPS. These features are then fed into the trained XGBoost Classifier to make a prediction regarding the website's legitimacy.

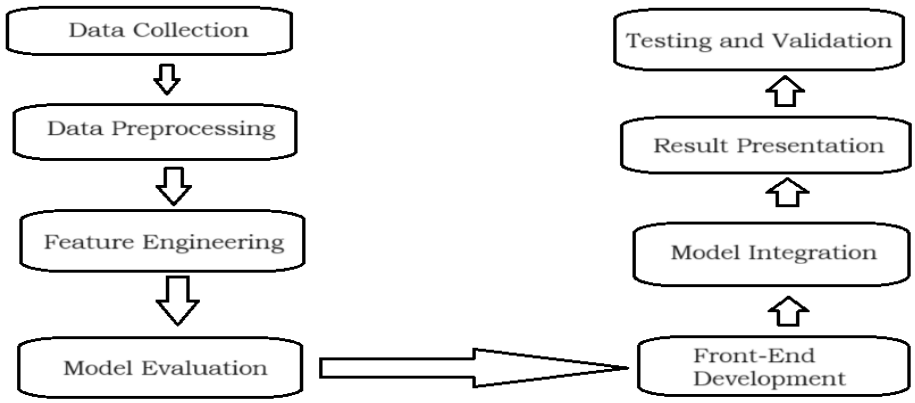


Figure 1: Methodology of Proposed Method

This methodology ensures a streamlined approach to phishing detection, leveraging machine learning algorithms and feature extraction techniques to analyze website attributes and make informed predictions. By focusing on the XGBoost Classifier, the methodology prioritizes accuracy and efficiency, culminating in a user-friendly front-end application capable of swiftly identifying potential phishing threats.

IV. RESULTS

This user-friendly application is designed to help individuals differentiate between phishing and legitimate websites. The app features clear and intuitive interfaces that guide users through the process of analyzing website authenticity. Sample outputs from the application are presented in the images below, offering users a visual representation of how the application operates and the results it delivers.

These images showcase the application's functionality, demonstrating how users can input a website URL and receive immediate feedback on whether the website is deemed phishing or legitimate. The application employs advanced algorithms to analyze key features of the website and make informed judgments based on predefined criteria. This process ensures that users can make informed decisions about the websites they visit, helping to mitigate the risk of falling victim to online scams or fraudulent activities.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304129 |

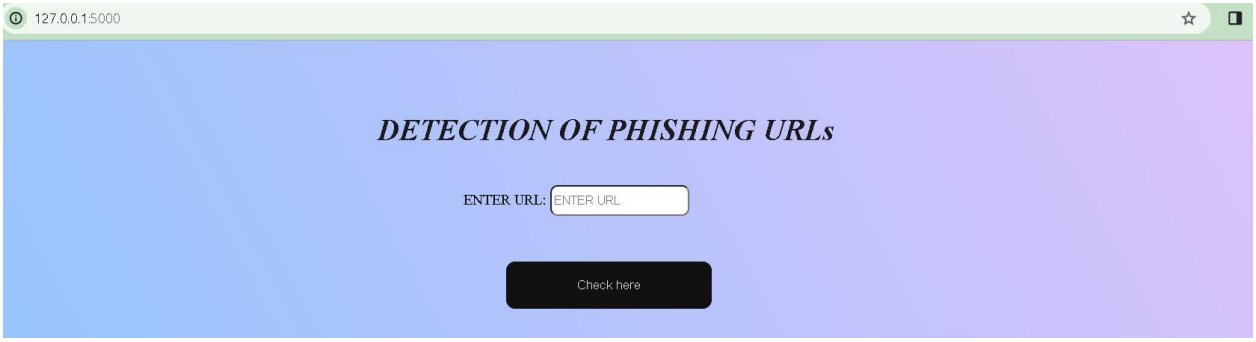
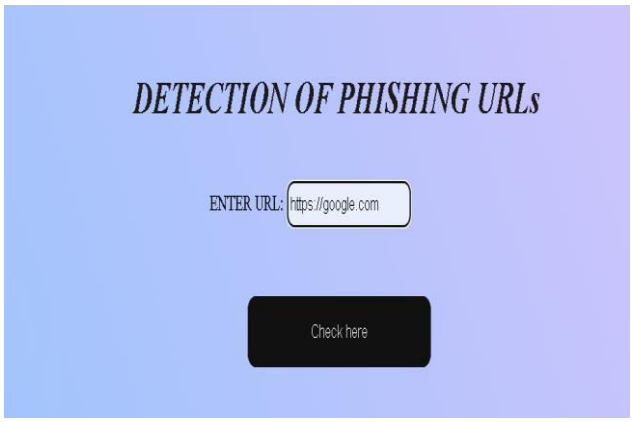


Figure 2: Application Interface of Proposed Method



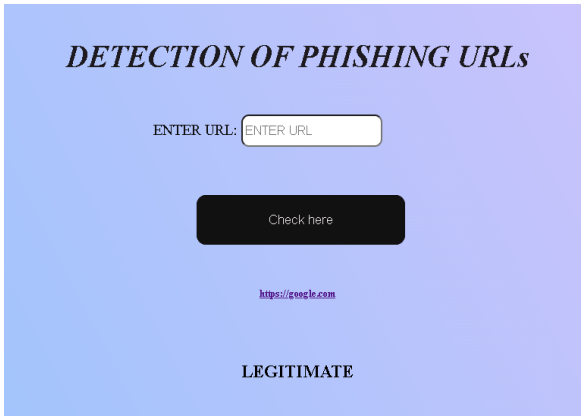


Figure 3: Legitimate URL Test Results

Figure 4: Phishing URL Test Results

V. CONCLUSION

In conclusion, our project on detecting phishing websites represents a significant step forward in the realm of cybersecurity, offering a proactive solution to combat the pervasive threat of online scams and fraudulent activities. Through the implementation of advanced machine learning algorithms and feature extraction techniques, we have developed an effective system capable of distinguishing between phishing and legitimate websites with high accuracy.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 4, April 2024 ||

| DOI:10.15680/IJIRSET.2024.1304129 |

By leveraging these technological advancements, our project not only enhances online security measures but also empowers users to make informed decisions about the websites they interact with, thereby mitigating the risks associated with cyber threats. Moving forward, our findings underscore the importance of continued research and development in the field of cybersecurity, as we strive to stay one step ahead of cybercriminals and safeguard the digital landscape for all users.

REFERENCES

1. Alswailem, A., Alabdullah, B., Alrumayh, N., & Alsedrani, A. (2019). Detecting Phishing Websites Using Machine Learning. 2019 2nd International Conference on Computer Applications & Information Security (ICCAIS).
2. Razaque, A., Frej, M. B. H., Sabyrov, D., Shaikhyn, A., Amsaad, F., & Oun, A. (2020). Detection of Phishing Websites using Machine Learning. 2020 *IEEE* Cloud Summit.
3. Alkawaz, M. H., Steven, S. J., & Hajamydeen, A. I. (2020). Detecting Phishing Websites Using Machine Learning. 2020 16th IEEE International Colloquium on Signal Processing & Its Applications (CSPA).
4. https://www.kaggle.com/datasets/eswarchandt/phishing-website-detector

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details