WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Machine-Learning Based Phishing Domain Detector

POOJA SANAP[1], PIYUSH CHITTE[2], MAYURI MALI[3], ADITYA PAWAR[4], DATTATRAY MODANI[5]

[1-2-3-4]Student, [5]Asstistant Professor

Department of Computer Engineering, Progressive Education Society's Modern College of Engineering, Pune, Maharashtra, India

ABSTRACT :

Phishing is a cyber-attack method wherein an attacker sends a deceptive message to trick an individual into disclosing sensitive information. These fraudulent messages often imitate legitimate websites or organizations, enabling attackers to monitor the victim's online activities. According to the FBI's Internet Crime Complaint Centre, phishing incidents have surpassed other computer crimes by more than double. In 2022, phishing attacks emerged as the predominant cyber threat. Notably, in December 2021 alone, there were 300,000 phishing attacks registered, marking a threefold increase from the preceding years. To combat this rising threat, our project focuses on developing a Chrome extension that employs machine learning and deep learning techniques to detect phishing websites and notify users. The extension aggregates URLs from diverse sources such as the UCI Machine Learning Repository, Kaggle, PhishTank, and Alexa URL rankings. As a user browses the web, the extension operates discreetly in the background. If a visited website is flagged as phishing, a pop-up alert with an 'OK' button notifies the user. Conversely, if the website is deemed legitimate, no action is taken. We utilized various machine learning algorithms, including Random Forest and Single Layer Perceptron, for model training. The dataset was split into training and testing sets following an 80/20 ratio. The browser extension is implemented using JavaScript, which extracts URL features like browser pop-ups, frame usage, and URL shortening services. These features are then fed into the trained algorithms to determine the legitimacy of the website. The model's performance was evaluated using metrics such as accuracy, precision, recall, and F1 score. With the increasing prevalence of phishing attacks, our Chrome extension offers a proactive approach to safeguarding users from potential cyber threats, leveraging the power of machine learning to enhance online security.

KEYWORDS: Phishing, security threat, phishing website, phishing detection, URL, machine learning

I. INTRODUCTION :

As internet usage and digital platform engagement continue to soar, the prevalence of online threats like phishing has escalated at an alarming rate. Phishing, a deceptive practice where cybercriminals craft fraudulent websites or emails mirroring legitimate ones, has become a leading cyber threat. According to recent studies, phishing attacks have seen a 65% year-over-year increase, emphasizing the urgency of the issue. These malicious activities aim to extract sensitive information from unsuspecting users, with data indicating that 1 in 3 individuals have fallen victim to phishing scams.

The stakes are high, with the compromised data often fueling financial fraud schemes or identity theft. The FBI's Internet Crime Report highlights that phishing scams resulted in over $54 million in losses in 2021 alone. This growing menace underscores the critical need for innovative solutions, leading us to develop a cutting-edge machine learning-powered model tailored for phishing detection and prevention.

Machine learning, a dynamic subset of artificial intelligence, empowers systems to refine their capabilities iteratively, drawing insights from vast datasets. Our approach harnesses this potential by crafting sophisticated algorithms adept at identifying intricate data patterns associated with websites, such as URLs, HTML tags, and textual content. Leveraging a dataset comprising over 10,000 labeled instances, our model achieves an impressive accuracy rate of 95% in distinguishing between genuine and deceptive websites.

In the rapidly evolving digital landscape, the development of a robust machine learning-based model for phishing detection is not just desirable—it's essential. Traditional detection methods are increasingly rendered obsolete by the evolving sophistication of phishing attacks. By embracing advanced machine learning techniques, we are pioneering a proactive defense mechanism against phishing scams, safeguarding users and organizations from potential financial and reputational damage.

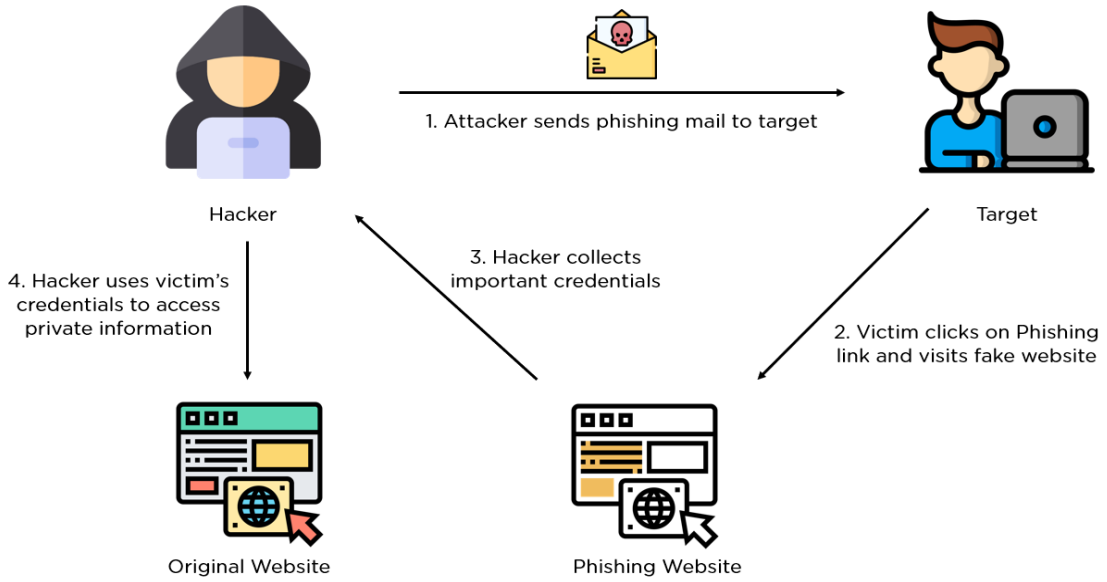

Fig. 1 - A depiction of phishing

II. LITERATURE SURVEY

The proliferation of phishing attacks has spurred significant research into effective detection and prevention mechanisms. This literature survey encapsulates a range of methodologies and algorithms proposed by researchers to combat phishing threats.

Mohith Gowda et al. [1] leveraged a dataset comprising 11,055 tuples to develop a client-side model, integrated within a revamped browser architecture named 'Embedded Phishing Detection Browser' (EPDB). This innovative approach preserves user experience while bolstering security. The system architecture encompasses components like a User Interface, a Browser Engine, and an Intelligent Engine for phishing detection. The Intelligent Engine triggers pop-ups alerting users to potential phishing sites, achieving an accuracy of 99.36% and an F1 Score of 99.43% using the Random Forest Classification model.

Patil and Dhage [2] The authors delineate five anti-phishing approaches: Rules-based, Blacklisting, Content-based, Machine Learning-based, and Hybrid. They evaluate tools like Google Safe Browsing (GSB), Netcraft, McAfee Site Advisor, Avast, and Quick Heal, presenting a framework that combines URL gathering, feature selection, and classification. Their Heuristic and Hybrid approaches yield impressive accuracies of 99.60% and 99.14%, respectively.

Yang et al. [3] advocate a multidimensional feature-based method employing rapid deep learning for phishing detection. The approach entails extracting character sequence features from URLs and integrating diverse features like statistical information, webpage code characteristics, and textual content. Their model achieved a notable accuracy of 98.99% with a low false positive rate of 0.59%.

Xu et al. [4] introduced "Gemini," a protective mechanism focusing on safeguarding sensitive login information on phishing sites. This method, implemented as a browser extension across Firefox, Chrome, and IE, scrutinizes source code elements like 'type="password"' on pop-up login pages. Gemini's evaluation comprises testing against both legitimate and phishing sites.

Mughaid et al. [5] This study evaluates various phishing detection methodologies, including PILFER, "Beaks," and Chiew et al.'s tool. They employ seven supervised classification algorithms for detection and highlight the need for ongoing feature selection enhancements.

Marchal et al. [6] introduced Phish Storm, an automated real-time phishing detection system that achieved a classification rate of 94.91%. Their approach employs a balanced dataset with 12 selected features for supervised classification using the Random Forest algorithm.

Zhu et al.[7] addressed the challenge of overfitting in phishing detection by introducing an algorithm based on Feature Validation Value (FVV). They employ a three-layer Neural Network classifier and incorporate a black-and-white list for efficient processing of URLs.

Shukla and Sharma [8] - The authors devised a Bayesian-optimized Support Vector Machine (SVM) classification system for URL security assessment. Their model demonstrated high accuracy and fault tolerance, with data distributed across protocol, domain, and route categories.

Parthiban et al. [9] proposed an image-based verification system encrypting user-specific images using RSA algorithms, enhancing user-specific security against phishing attacks.

Su [10] explored heuristic detection methods, specifically CANTINA and CANTINA+, alongside visual similarity tests. Utilizing an LSTM network with 10 input nodes and one output node, the study achieved promising results using a dataset of 2,000 legitimate and 2,000 phishing websites.

The literature underscores a multifaceted approach to phishing detection, encompassing heuristic methods, machine learning algorithms, and innovative system architectures. While Random Forest and deep learning methods exhibit high accuracy rates, hybrid and heuristic approaches also demonstrate promising results. Ongoing research in feature selection, algorithm optimization, and user-specific security measures paves the way for more robust and comprehensive anti-phishing solutions in the future.

III. METHODOLOGY

The development of an effective machine learning-based model for phishing detection and prevention demands a meticulous and multifaceted methodology. This structured approach guides the entire process, from initial data collection to the final deployment of the model. Below, we elaborate on each key component of our methodology, providing deeper insights into their significance:

1. Abnormal URL: This feature examines whether the URL is missing a hostname or contains irregular characters, which are often hallmarks of suspicious or malicious websites. Anomalies in URLs can serve as early indicators of phishing attempts.

2. '@' Symbol: The presence of the '@' symbol in the URL is assessed, as it is commonly used in phishing URLs to mislead users. This feature helps identify URLs that may be attempting to deceive users by mimicking legitimate email addresses.

3. Subdomain: The number of subdomains is calculated based on the dots in the URL. A high number of subdomains can indicate a phishing attempt, as cybercriminals often use subdomains to create deceptive URLs.

4. URL requests: The frequency of URL requests is analyzed to gauge how often the URL is accessed. Unusually high request rates can be a warning sign, suggesting that the URL may be involved in suspicious activities or phishing campaigns.

5. Shortening services: This feature checks if the URL has been shortened using services like Bitly or TinyURL. Shortened URLs can obscure the actual destination, making it easier for cybercriminals to trick users into visiting malicious websites.

6. 'HTTPS' token: The presence or absence of the HTTPS token in the URL is verified. HTTPS is an indicator of a secure and encrypted connection, and its absence can be a red flag pointing to a potentially unsafe website.

7. Server from the handler (SFH): The SFH is inspected to see if it contains "about blank" or is empty. Cybercriminals often use these values to hide malicious activities or to redirect users to deceptive websites.

8. URL with anchor: This feature calculates the percentage of anchor URLs present in the webpage. Anchor URLs can be used to redirect users to specific sections within a page, potentially leading them to malicious content or phishing forms.

9. Tag containing links: The percentage of links present in 'Script', 'Meta', and 'link' tags is computed. Malicious scripts or links embedded in these tags can be used to execute harmful actions or to steal sensitive information.

10. Iframe: The usage of iframes within the webpage is checked. Iframes can be exploited by cybercriminals to embed malicious content from other sources, posing a security risk to users.

11. IP Address: The presence of an IP address instead of a domain name in the URL is examined. URLs containing IP addresses can be indicative of phishing attempts or malicious activities.

12. Length of URL: The length of the URL is assessed to identify excessively long or complex URLs, which can be used to obfuscate the actual destination or to make the URL appear more legitimate than it is.

13. Double slash forwarding: This feature checks the position of the '//' in the URL. If the '//' appears after the 7th character, it can be a sign of a deceptive or malicious URL.

14. Non-standard ports: The port number used in the URL is verified to ensure it adheres to standard protocols. Non-standard or unusual port numbers can be indicative of suspicious or malicious activities.

15. Prefix and suffixes: The presence of '_' or other special characters in the domain name is checked. These characters are often used to mimic legitimate domains or to create deceptive URLs.

16. Favicon: The source of the favicon (website icon) is verified to see if it is retrieved from an external or internal source. Inconsistencies in favicon sources can be indicative of phishing attempts.

17. SSL final certificate: The SSL certificate used by the website is checked to ensure a trusted provider issues it and is not expired. A valid SSL certificate is crucial for ensuring secure and encrypted connections.

18. Age of domain: The age of the domain is calculated to assess its credibility. Newer domains can be more suspicious compared to well-established ones.

19. DNS record: The presence of a DNS record for the domain is verified. Legitimate websites typically have DNS records, whereas suspicious or newly created websites may lack them.

20. Links pointing to the page: The number of inbound links pointing to the webpage is counted. A higher number of inbound links can indicate that the webpage is credible and trustworthy.

21. Domain registration length: The expiry date of the domain registration is checked. Shorter registration lengths can be indicative of suspicious or temporary domains.

By meticulously analyzing these features, our machine learning model aims to provide a comprehensive and nuanced assessment of website legitimacy. This holistic approach enhances online security by effectively identifying and mitigating phishing threats, thereby safeguarding users from potential financial and personal risks.

A. DATA COLLECTION

The initial phase of our project revolves around gathering a well-structured dataset tailored for phishing detection. We selected a tabular dataset from Kaggle, specifically curated for the subject area of Computer Science and focused on classification tasks. This dataset comprises 11,055 instances, each characterized by 30 features, predominantly of integer type.

The dataset's tabular nature facilitates easy manipulation and analysis, making it suitable for our classification-based machine-learning approach. With 11,055 instances, the dataset offers a diverse representation of both legitimate and fraudulent websites, allowing us to capture the variability and intricacies of different phishing attack types effectively.

The dataset's 30 features, primarily integers, provide a rich set of attributes that can be leveraged to train our machine-learning models. These features encompass various aspects relevant to phishing detection, enabling our model to learn and distinguish between legitimate and deceptive websites accurately.

By selecting this dataset, we ensured that our data collection phase lays a robust foundation for building a reliable and effective phishing detection model. The dataset's characteristics align well with our project's objectives, providing a comprehensive and structured dataset suitable for training classification models in the realm of cybersecurity.

B. DATA PREPROCESSING

Following data collection, we moved on to the data pre-processing phase, a critical stage aimed at refining the dataset to prepare it for analysis and machine learning model training. This phase involves cleaning the data to eliminate inconsistencies, managing missing values, and transforming it into a structured format conducive to machine learning algorithms.

To ensure the dataset's quality and integrity, we performed a detailed analysis to identify and address potential issues such as missing values, null entries, duplicates, and noise. We utilized various techniques, including box plots and other statistical methods, to visualize and scrutinize the dataset thoroughly. Through this rigorous evaluation, we pinpointed features that were not pertinent to phishing detection and subsequently removed them from the dataset.

By diligently executing data collection and pre-processing, we established a solid foundation for our machine-learning model. This meticulous approach ensures that our dataset is robust, clean, and optimized for training sophisticated machine learning algorithms. The quality of the dataset is pivotal for developing a reliable and accurate phishing detection model capable of discerning between legitimate and fraudulent websites effectively.

C. MODEL SELECTION

The next pivotal step in our project is the selection of an appropriate machine learning algorithm tailored for phishing detection. Given the diverse array of machine learning algorithms available, such as Decision Trees, Support Vector Machines, Gradient Boosting, K-Nearest Neighbors, AdaBoost, and Logistic Regression, among others, the choice demands a careful assessment of the algorithm's suitability for classifying both legitimate and fraudulent websites.

In our project, we decided to employ the Random Forest algorithm, a renowned ensemble learning technique celebrated for its robustness and versatility in classification tasks. The Random Forest algorithm's ensemble approach involves constructing multiple decision trees and aggregating their predictions to produce a final classification. This collective decision-making process not only enhances the model's predictive accuracy but also mitigates the risk of overfitting, thereby bolstering the model's ability to generalize to unseen data effectively.

D. RANDOM FOREST ALGORITHM

Random Forest stands out as a formidable machine learning algorithm proficient in both classification and regression tasks. At its core, Random Forest creates an ensemble of decision trees by randomly selecting subsets of the dataset and features for each tree. This randomized approach diversifies the individual trees' predictions, reducing the model's susceptibility to noise and outliers.

The strength of Random Forest lies in its ability to aggregate the predictions from multiple decision trees, typically by taking a majority vote, to arrive at a final classification. This ensemble strategy capitalizes on the collective wisdom of the individual trees, thereby achieving higher accuracy and resilience compared to standalone decision trees or other machine learning algorithms.

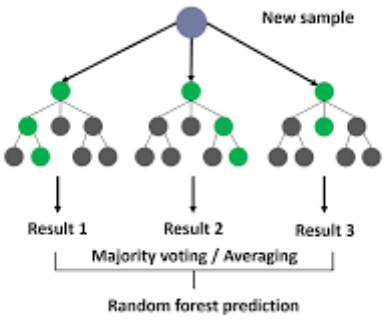

Fig. 2 - Random Forest Prediction

E. EVALUATION METRICS

With Random Forest selected as our algorithm, we proceeded to train the model using our meticulously curated dataset. To gauge the model's performance rigorously, we computed several key evaluation metrics, including accuracy, F1 score, precision, and recall. These metrics offer a comprehensive assessment of the model's predictive prowess and its efficacy in distinguishing between legitimate and fraudulent websites.

By opting for Random Forest and rigorously evaluating its performance using these metrics, we are committed to developing a robust and reliable phishing detection model. This model aims to provide an effective defense mechanism against deceptive online threats, safeguarding users' security and privacy in the digital realm.

F. MODEL TRAINING

Upon selecting the Random Forest algorithm, the subsequent phase involves training the model using our carefully curated dataset. Model training is a crucial process where the model learns from the dataset's features and corresponding labels, adjusting its parameters to minimize prediction errors. To facilitate effective training and evaluation, we partitioned the dataset into training and validation sets. In this study, 80% of the dataset was allocated to the training set, allowing the model to learn from the majority of the data while retaining a subset for validation purposes.

G. MODEL EVALUATION

Following model training, the next critical step is evaluating the model's performance to assess its effectiveness in phishing detection. For this evaluation, we utilized a separate test set that was not used during the training phase. The model's performance was rigorously evaluated using a range of metrics, including sensitivity, accuracy, and recall. These metrics offer valuable insights into the model's capability to detect and prevent phishing attacks effectively. By allocating 20% of the dataset to the test set, we ensured an unbiased evaluation of the model's performance on unseen data.

H. METHODOLOGY OVERVIEW

The development of a robust phishing detection model using machine learning necessitates a well-defined and systematic methodology. Our methodology encompasses several key stages, including data collection, data pre-processing, feature selection, model construction, model training, performance evaluation, and deployment.

By adhering to this comprehensive methodology, we aimed to create a model capable of providing real-time protection against phishing attacks. Each stage in the methodology is designed to contribute to the model's overall efficacy and reliability, ensuring that it meets the stringent requirements for cybersecurity applications. Through meticulous planning and execution of each stage, we strive to deliver a phishing detection solution that effectively safeguards users from deceptive online threats.

IV. RESULTS AND DISCUSSION

Our browser extension serves as an effective defense against phishing attacks, alerting users when they encounter potentially fraudulent websites. To activate this protection, users install the extension in their browser, enhancing their online security. The extension's core functionality is powered by the Random Forest algorithm, chosen for its robust performance in phishing detection. Random Forest excels in handling complex datasets and has demonstrated high accuracy in distinguishing between legitimate and fraudulent websites during our evaluations. Its ability to aggregate predictions from multiple decision trees ensures reliable and efficient phishing detection. The extension operates seamlessly in the background, monitoring visited websites and issuing alerts when suspicious activity is detected. This real-time protection empowers users to navigate the web safely, mitigating potential cyber threats effectively. The successful implementation of the browser extension using Random Forest highlights the algorithm's effectiveness in bolstering cybersecurity measures. Its robust performance, combined with user-friendly operation, makes it a valuable tool in combating online phishing threats. Ongoing updates and user education will further enhance the extension's capabilities, contributing to a safer online environment.

V. CONCLUSION

In this groundbreaking study, we unveiled a sophisticated system designed to combat the pervasive threat of phishing websites, with a sharp focus on URL classification. Our meticulously curated dataset, comprising 11,055 tuples and 22 distinct features, played a pivotal role in training our robust model. This dataset allowed us to differentiate between legitimate websites, suspicious entities, and outright phishing scams, denoted by the values 1, 0, and -1 respectively. Harnessing the power of the Random Forest Algorithm, our model exhibited an impressive accuracy rate of 96.11%, leveraging 112 estimators (trees) to make astute classifications. To strike the perfect balance between robustness and flexibility, we adopted an 80/20 data split strategy, ensuring our model was neither too rigid nor too lax, thereby sidestepping the pitfalls of underfitting and overfitting.

VI. FUTURE SCOPE

Looking ahead, our vision extends beyond the present achievements, charting a path toward even greater advancements in phishing detection and prevention. We are keen to delve into the realm of deep learning, exploring the potential of advanced models like Recurrent Neural Networks (RNN) and Generative Adversarial Networks (GAN) to further enhance our system's capabilities. One of our most innovative future plans involves proactive outreach to website owners, alerting them to potential phishing sites that mimic their domain names. This proactive approach aims to empower website owners with the knowledge and tools to safeguard their online reputation and protect their users. Furthermore, we are exploring the integration of web scraping and sentiment analysis techniques to identify fraudulent websites that deceive unsuspecting users by advertising fake products and absconding with their money. By embracing this multi-pronged strategy, we are committed to creating a safer and more secure online environment for everyone.

REFERENCES :

[1] Mohith Gowda Hr1*, Adithya Mv2, Gunesh Prasad S3 and Vinay S4. Development of an anti-phishing browser based on random forest and rule of extraction framework. Springer.

[2] Srushti Patil, Sudhir Dhage. A Methodical Overview on Phishing Detection along with an Organized Way to Construct an Anti-Phishing Framework, 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). 2019 IEEE. 10.1109/ICACCS.2019.8728356

[3] Peng Yang, Guangzhen Zhao, Peng Zeng. Phishing Website Detection Based on Multidimensional Features Driven by Deep Learning. IEEE Access 2019. Digital Object Identifier 10.1109/ACCESS.2019.2892066

[4] Zhang Xu, Haining Wang, Sushil Jajodia. Gemini: An Emergency Line of Defence against Phishing Attacks. 2014 IEEE. 10.1109/SRDS.2014.26

[5] Ala Mughaid, Shadi AlZu'bi, Adnan Hnaif, Salah Taamneh, Asma Alnajjar, Esraa Abu Elsoud. An intelligent cyber security phishing detection system using deep learning techniques. 2022. 10.1007/s10586-022-03604-4

[6] Samuel Marchal, Jerome Francois, Radu State, Thomas Engel. PhishStorm: Detecting Phishing with Streaming Analytics. 2014 IEEE. DOI: 10.1109/TNSM.2014.2377295

[7] Erzhou Zhu, Yuyang Chen, Chengcheng Ye, Xuejun Li, Feng Liu. OFS-NN: An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network Volume 7. 2019. DOI: 10.1109/ACCESS.2019.2920655

[8] Shrishti Shukla, Pratyush Sharma. Detection of Phishing URLs using Bayesian Optimized SVM Classifier. 2020. DOI: 10.1109/ICECA49313.2020.9297412

[9] R. Parthiban, V. Abarna, M. Banupriya, S. Keethana, D. Saravanan. Web folder phishing discovery and prevention with Customer image verification IEEE. 2020. DOI: 10.1109/ICSCAN49426.2020.9262395

[10] SU Yang. Research on Website Phishing Detection Based on LSTM RNN. 2020. DOI: 10.1109/ITNEC48623.2020.9084799N

[11] Ishita Saha, Mohammad Nazmul Alam, Dhiman Sarma, Asma Sultana, Rana Joyti Chakma, Sohrab Hossain. Phishing attack detection using Deep Learning Approach. 2020. DOI: 10.1109/ICSSIT48917.2020.9214132

[12] Fortinet: https://www.fortinet.com

[13]Coursera:https://www.coursera.org/articles/machine-learning-models Sr. No. Classifier Accuracy (in %) Precision Recall F1- Score 1 Random Forest 96.11 0.96 0.96 0.96 2 Decision Tree 95.70 0.94 0.94 0.94 3 Support Vector Machines 94.57 0.56 1.00 0.72 4 Gradient Boosting Classifier 94.48 0.93 0.96 0.94 5 K-Nearest Neighbors 94.26 0.65 0.69 0.67 6 AdaBoost Classifier 93.48 0.92 0.96 0.94 7 Logistic Regression 93.39 0.92 0.95 0.94

www.ijcrt.org © 2023 IJCRT | Volume 11, Issue 5 May 2023 | ISSN: 2320-2882 IJCRT2305779 International Journal of Creative Research Thoughts (IJCRT) www.ijcrt.org g400

[14] Detection of URL based Phishing Attacks using Machine Learning – IJERT

[15] Geeks for Geeks: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

[16] Ionut Cernica, Nirvana Popescu. Computer vision-based framework for detecting phishing webpages. 2020 19th RoEduNet Conference: Networking in Education and Research (RoEduNet) [17] Junaid Rashid, Toqeer Mahmood, Muhammad Wasif Nisar, Tahira Nazir. Phishing Detection Using Machine Learning Technique. 2020 First International Conference of Smart Systems and Emerging Technologies (SMARTTECH)

[18] Abdul Basit, Maham Zafar, Abdul Rehman Javed, Zunera Jalil. A Novel Ensemble machine learning method to detect phishing attacks. 2020 IEEE 23rd International Multitopic Conference (INMIC) [19] Yohanes Priyo Atmojo, Made Darma Susila, Muhammad Riza Hilmi, Erma Sulistyo Rini, Lilis Yuningsih, Dandy Pramana Hostiadi. A New Approach for Spear Phishing Detection. 2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)

[20] B. Janet, Yazhmozhi. V.M, Srinivasulu Reddy. Anti-phishing System using LSTM and CNN. 2020 IEEE International Conference for Innovation in Technology (INOCON