PHISHING URL DETECTION A REAL-CASE SCENARIO THROUGH LOGIN URLS

Mrs N. Bhargavi[1], V. Sravani[2], K. Akshaya Sree[3]
[1]Assistant professor, Department of CSE, Princeton College of engineering and technology for women Narapally vijayapuri colony ghatkesar mandal, Pin code-500088
[2,3]UG Students, Department of CSE, Princeton College of engineering and technology for women Narapally vijayapuri colony ghatkesar mandal, Pin code-500088

ABSTRACT:

Phishing attacks, where malicious actors attempt to deceive users into divulging sensitive information through fake websites, have become a significant cybersecurity threat. One common tactic is to create fake login pages that mimic legitimate websites to steal user credentials. This study focuses on the detection of phishing URLs in real-case scenarios, specifically through the analysis of login URLs. Phishing attacks have evolved to become more sophisticated, making it challenging for users to distinguish between legitimate and fake websites. Phishers often use login pages of popular services, such as banking, email, and social media platforms, to lure victims into revealing their login credentials. Detecting these phishing URLs is crucial to protecting users from identity theft, financial loss, and unauthorized data access. In this research, a methodology is proposed to detect phishing URLs in real-case scenarios, with a specific emphasis on login URLs. The methodology combines machine learning techniques, web content analysis, and URL characteristics assessment. Features such as domain similarity, SSL certificate validity, page content analysis, and URL structure are utilized to determine the likelihood of a URL being a phishing attempt.

Keywords: URL, SSL, phishing attacks, SVM, dataset.

I INTRODUCTION

Phishing attacks represent a persistent and significant cybersecurity threat that continues to exploit human vulnerabilities to compromise sensitive information, credentials, and financial assets. These attacks employ various social engineering tactics to deceive users into divulging their personal and confidential data. One common and

effective approach employed by phishers is to create deceptive login URLs, mimicking legitimate websites to lure users into providing their login credentials. In recent years, the proliferation of online services and the increasing reliance on digital platforms have amplified the impact of phishing attacks. The primary goal of phishing is to manipulate users into unknowingly sharing their login credentials, which can then be exploited by cyber criminals for financial gain or unauthorized access to sensitive information. The use of login URLs in phishing attacks is particularly concerning due to their direct association with users' trust and reliance on authentic online services. These attacks capitalize on users' expectations of entering their credentials on familiar login pages, leading them to overlook potential discrepancies that might indicate a fraudulent website. Traditional phishing detection methods often struggle to keep pace with the evolving techniques used by attackers, making it crucial to develop advanced approaches that can effectively identify phishing URLs in real-case scenarios.

This research focuses on addressing the challenges posed by phishing attacks that employ deceptive login URLs. By analyzing real-case scenarios, this study aims to develop and evaluate a comprehensive phishing URL detection methodology that enhances the ability to identify these fraudulent URLs accurately. Leveraging machine learning techniques, web content analysis, and URL attributes, this methodology seeks to provide a robust defense against the growing threat of phishing attacks. In the subsequent sections, we will delve into the specific methodologies employed, the datasets used for training and evaluation, the results obtained, and the implications of the findings. By advancing the state of phishing detection technology, this research contributes to the broader effort of enhancing online security and protecting users from falling victim to phishing attacks facilitated through deceptive login URLs.

Problem statement:

Phishing attacks remain a persistent and evolving threat to cybersecurity, exploiting human psychology and technological vulnerabilities to compromise sensitive information. A particularly potent variant of these attacks involves the creation of deceptive login URLs, which mimic legitimate websites to deceive users into revealing their login credentials. Detecting and thwarting these attacks is a critical challenge in

safeguarding users' personal data and digital assets.

In the context of real-case scenarios, the problem lies in the difficulty of accurately identifying phishing URLs, especially those that leverage login pages. Traditional anti-phishing methods often rely on known patterns and heuristics, which attackers can easily evade by employing new tactics and exploiting users' trust in familiar login interfaces. This challenge is exacerbated by the rapid proliferation of online services and the use of sophisticated obfuscation techniques by cybercriminals.

The goal of this research is to develop an effective and robust methodology for detecting phishing URLs, with a specific focus on those utilizing deceptive login pages, within the context of real-case scenarios. This methodology aims to overcome the limitations of existing detection methods by combining advanced techniques such as machine learning, web content analysis, and URL characteristics assessment. The key issues to address include:

Evolving Tactics: Attackers continually modify their techniques to bypass traditional detection methods. The proposed methodology must be capable of identifying new and sophisticated phishing URL variants that leverage deceptive login pages.

User Trust Exploitation: Phishers capitalize on users' implicit trust in familiar login interfaces. The challenge is to develop techniques that uncover even subtle deviations from genuine login URLs that users might overlook.

Real-Case Context: Phishing attacks occur within dynamic and diverse environments. The methodology should be capable of analyzing URLs in real-time, across various industries and services, to ensure its practical applicability.

Balancing False Positives: Striking a balance between accurate detection and minimizing false positives is crucial. The methodology must achieve high precision to prevent genuine URLs from being incorrectly flagged as phishing attempts.

Adaptability: As attackers adapt and change their strategies, the detection methodology needs to be flexible and adaptable to emerging threats without requiring frequent manual updates.

Addressing these challenges requires a multidimensional approach that leverages the latest advancements in cybersecurity, machine learning, and web analysis. By doing so, this research

aims to contribute to the development of effective solutions that enhance the security posture of online users, organizations, and digital ecosystems against the pervasive threat of phishing attacks through deceptive login URLs.

II LITERATURE SURVEY

Rashmi Karnik et al., proposed a model of classification method, kernel-based approach. In this we categories phishing . This method produces estimated accuracy of 95% in detecting the phishing and malware sites.

Andrei Butnaru et al., used a supervised Machine Learning algorithm to block phishing attacks, based on novel mixture phishing attacks and compare with Google Safe browsers.

Vahid Shahrivari et al., proposed a one of the most successful techniques for identifying these malicious works is Machine Learning. It is because of most Phishing attacks have same features which can be noticed by Machine learning techniques. In this many machine learning-based classifiers are used for forecasting the phishing websites. The main advantage of machine learning is the ability to create flexible models for specific tasks like phishing detection. Since phishing is a classification problem, Machine learning models can be used as a forceful tool.

Ammara Zamir et al., proposed a framework for identifying phishing websites using a heaping model. Information gain, gain ratio, Relief-F, and recursive feature elimination (RFE) are some of the feature selection algorithms that can be used to analyse Phishing characteristics. The greatest and weakest traits are combined to create two features. Bagging is used in principal component analysis using several Machine learning algorithms, including random forest [RF] and neural network [NN]. Two heaping representations heaping1 (RF + NN + Bagging) and heaping2 (kNN + RF + Bagging) are applied by merging highest scoring classifiers to progress classification accuracy.

Nguyet Quang Do, Ali Selamat et al., conducted a study on phishing detection and proposed a four different deep learning technique, includes deep neural network (DNN), convolution neural networks (CNN), Long Short-term memory (LSTM), and gated recurrent unit (GRU). To analyse behaviour of these deep learning architectures, extensive experiments were carried out to examine the impact of parameter tuning on the performance accuracy of

the deep learning models. In which each model shows different accuracy's from different models.

Ashit Kumar Dutta proposed a URL detection procedure based on Machine Learning methods. An RNN is used for identifying the phishing URL. It is evaluated with 7900 malicious and 5800 genuine sites, respectively. The outcome of this method shows a good concert compare to recent tactics.

Atharva Deshpande et al., proposed a combination of machine learning algorithms and natural language processing methods to detect the phishing domain appearances, the feature that distinguish them from real domains. Ms.

Sophiya Shikalgar et al., proposed a machine learning classifiers and methods to detect phishing website using Hybrid machine learning approach is a combination of different classifiers working together which gives a good prediction result. Each of classifiers have its own way of working and classification. Uses a data of URLs which contains 2905 URLs which is in unstructured form.

Nureni Ayofe Azeez et al., tried to handle this challenge, attempts have been made to address two major problems. The first is how can the suspicious URL's be recognized on social networks and how can internet users can be protected from unreliable and fake URLs on the social network. It adapts six machine learning methods – AdaBoost, Gradient Boost, random forest, Linear SVM, decision tree and Naïve Bayes classifier for training using features obtained from the social network and for additional processing. A total of 532,403 posts were analysed. At last 87,083 posts were considered suitable for training the models. AdaBoost performs well among all with an accuracy of 95% and a precision of 97%. Ademola Philip Abidoye and Boniface Kabaso proposed a machine learning technique to accurately classify the dataset to identify the phishing URLs features that can be used by the attackers.

R. Kiruthiga and D. Akila explained a novel way of detecting phishing websites using machine learning methods and proposes a classification model in order to classify the phishing attacks. Also presents a way to detect phishing email attacks using natural language processing and machine learning produces a good accuracy.

III. EXISTING SYSTEM

Existing solutions detect mimicked phishing pages by either text-based features or visual similarities of webpages and it can be easily bypassed and proposed a technique to identify the real domain name of a visiting webpage based on signatures created for web sites, site signatures, including distinctive texts and images, can be generated by analyzing common parts from pages of a website. The authors claimed that the method achieves high accuracy and low error rates. Aaron Blum et. eexplored the possibility of utilizing confidence weighted classification combined with content-based phishing URL detection to produce a dynamic and extensible system for detection of present and emerging types of phishing domains, and authors further claim the system can detect emerging threats and can provide an increased protection against zero-hour threats, unlike traditional blacklisting techniques which function reactively.Exist in phishing attacks in reality and can detect zero-hour phishing attack. But the characteristics are not guaranteed to always exist in such attacks and false positive rate in detection is very high. This tag is used to add another web page into existing main webpage. Phishers can make use of the “iframe” tag and make it invisible i.e. Without frame borders. Since border of inserted webpage is invisible, user seems that the inserted web page is also the part of the main web page and can enter sensitive information.

IV. PROPOSED SYSTEM

As we have moved most of our financial, work related and other daily activities to the internet, we are exposed to greater risks in the form of cyber crimes. URL based phishing attacks are one of the most common threats to the internet users. In this type of attack, the attacker exploits the human vulnerability rather than software flaws. It targets both individuals and organizations, induces them to click on URLs that look secure, and steal confidential information or inject malware on our system. Different machine learning algorithms are being used for the detection of phishing URLs, that is, to classify a URL as phishing or legitimate. Researchers are constantly trying to improve the performance of existing models and increase their accuracy. In this work we aim to review various machine learning methods used for this purpose, along with datasets and URL features used to train the machine learning models. The performance of different machine learning algorithms

and the methods used to increase their accuracy measures are discussed and analysed. The goal is to create a survey resource for researchers to learn the current developments in the field and contribute to making phishing detection models that yield more accurate results.

V.WORKING METHODOLOGY

Detecting phishing URLs in real-case scenarios through login URLs involves a combination of techniques and processes designed to identify fraudulent websites that imitate legitimate login pages. The methodology outlined below provides a high-level overview of the steps involved in effectively detecting phishing URLs targeting login credentials:

URL Collection and Preparation: Gather a diverse dataset of URLs from real-case scenarios. This dataset should encompass both legitimate and phishing URLs, covering a wide range of industries and services. Extract and prepare relevant features from the URLs, such as domain, subdomain, path, parameters, and protocol.

Feature Extraction: Extract additional features from the URLs, such as the length of the URL, the presence of hyphens or numbers, and domain reputation information. These features will be used to train and test machine learning models.

Machine Learning Model Training: Train machine learning models using the prepared dataset. Common algorithms include decision trees, random forests, support vector machines, and neural networks. Labels for the dataset should indicate whether each URL is legitimate or a phishing attempt.

Content Analysis: For suspected phishing URLs, fetch the content of the associated webpage. Analyze the webpage's structure, the presence of login forms, and the usage of external resources. These analyses can help determine the authenticity of the webpage.

SSL Certificate Verification: Check the validity of the SSL certificate associated with the URL. Verify if the SSL certificate matches the domain and is issued by a reputable certificate authority. An invalid certificate or a mismatch can indicate a phishing attempt.

Domain Reputation Check: Consult a database of known malicious domains to check if the URL matches any listed domains. If the URL is present in the database, it's likely a phishing attempt.

User Behavior Monitoring: Monitor user interactions with URLs, such as mouse movements, clicks, and typing patterns. Identify deviations from expected behavior on legitimate login pages, which could indicate phishing attempts.

Threshold Determination: Set appropriate thresholds for model predictions and feature analyses. These thresholds determine when a URL is flagged as suspicious or malicious. The thresholds should be balanced to minimize false positives and false negatives.

Real-Time Scanning and Reporting: Implement the detection methodology in a real-time environment. When users attempt to access URLs, apply the trained machine learning model and conduct the feature analyses. If a URL is identified as suspicious, provide warnings or block access, and report the incident.

User Education and Feedback: Educate users about phishing risks, best practices for identifying phishing URLs, and how to report suspicious URLs. Encourage users to provide feedback on flagged URLs to improve the system's accuracy over time.

Continuous Monitoring and Improvement: Regularly update the machine learning models and databases of known malicious domains. Stay informed about new phishing techniques and adapt the methodology to counter emerging threats.

The success of the methodology relies on a combination of advanced technological solutions, user awareness, and ongoing research to stay ahead of evolving phishing tactics. Continuous improvement, user feedback, and collaboration among cybersecurity professionals are key components in maintaining an effective phishing URL detection system within real-case scenarios involving login URLs.

VI.OPERATION:

Fig.1. Web URL page.

Phishing URL detection is an essential component in the ongoing battle to secure digital environments against cyber threats. As the internet continues to play an increasingly central role in our lives, the proliferation of phishing attacks poses a significant risk to users, organizations, and data integrity.

Detecting phishing URLs and preventing users from falling victim to fraudulent schemes is a critical endeavor that requires sophisticated technological solutions, user education, and continuous vigilance.

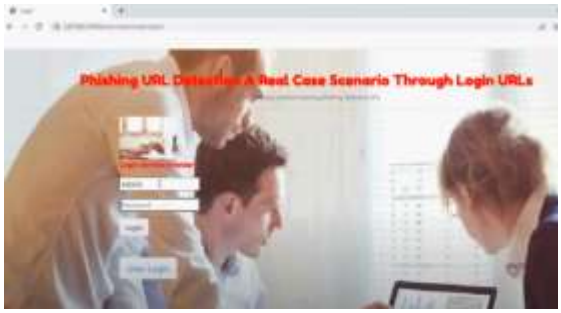

Fig.2. Admin login page.

Fig.3. User details.

Fig.4. Register details.

Fig.5. URL upload page.

In the pursuit of effective phishing URL detection, various approaches and methodologies have been developed, leveraging advancements in machine learning, web content analysis, user behavior monitoring, and domain reputation assessment. These techniques aim to identify malicious URLs that imitate legitimate websites, often exploiting users' trust and familiarity with common interfaces such as login pages.

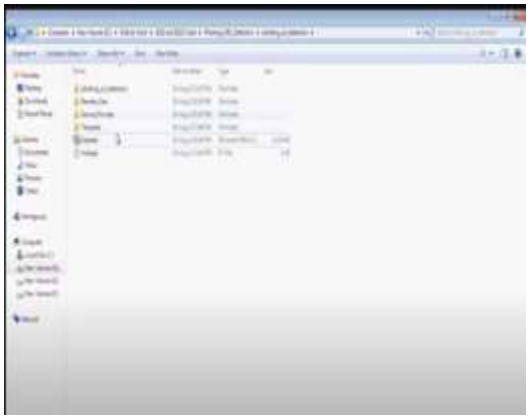

Fig.6. upload dataset details.

Fig.7. URL dataset with accuracy.

Despite the progress made in phishing URL detection, challenges persist. Cybercriminals continuously adapt their tactics to evade detection, necessitating the development of dynamic and adaptive defense mechanisms. The dynamic nature of the internet, the emergence of new attack vectors, and the increasing sophistication of phishing attacks underscore the need for ongoing research and innovation in this field.

User education remains a crucial aspect of combating phishing attacks. Empowering users with knowledge about the common signs of phishing, the importance of scrutinizing URLs, and the risks associated with sharing personal information can significantly contribute to a safer online experience.

Fig.8. Output Graphs.

Fig.9. Accuracy levels.

Fig.10. Phishing detected.

VII.CONCLUSION

In conclusion, the fight against phishing URL detection is an ongoing collaborative effort that involves researchers, cybersecurity professionals, technology providers, and users alike. As technology continues to evolve and

the threat landscape expands, staying ahead of cybercriminals requires a multidimensional strategy that combines advanced technological solutions with user education, awareness, and the commitment to fostering a secure digital ecosystem. By adopting a proactive and collaborative approach, we can work towards mitigating the risks posed by phishing attacks and building a more secure online environment for everyone.

VIII.REFERANCES

[1] Anti-phishing Working Group (APWG) Phishing Activity Trends Report 4th quarter 2020, https://docs.apwg.org/reports/apwg trends report q4 2020.pdf

[2] FBI Internet Crime Report 2020, https://www.ic3.gov/Media/PDF/Annual Report/2020 IC3Report.pdf

[3] Verizon 2020 Data Breach Investigation Report, https://enterprise.verizon.com/resources/ reports/2020-databreachinvestigations-report.pdf

[4] World Health Organization, Communicating for Health, Cyber Security, https://www.who.int/about/communications/cyber-security

[5] Ye Cao, Weili Han, and Yueran Le, "Anti-phishing based on automated individual white-list," Proceedings of the 4th ACM workshop on Digital identity management-DIM 08, pp. 51-60, 2008

[6] M. Sharifi, and S. H. Siadati, "A phishing sites blacklist generator," 2008 IEEE/ACS International Conference on Computer Systems and Applications, pp. 840-843, 2008

[7] N. Abdelhamid, A. Ayesh, and F. Thabtah, "Phishing detection based associative classification data mining," Expert Systems with Applications, vol. 41, no.13, pp. 5948-5959, 2014

[8] L. Wenyin, G. Huang, L. Xiaoyue, Z. Min, and X. Deng, "Detection of phishing webpages based on visual similarity," Special interest tracks and posters of the 14th international conference on World Wide WebWWW 05, pp. 1060-1061, 2005

[9] C. L. Tan, K. L. Chiew et al., "Phishing website detection using url assisted brand name weighting system," 2014 International Symposium on Intelligent Signal Processing and Communication Systems(ISPACS), IEEE, pp. 054-059, 2014

[10] K. L. Chiew, E. H. Chang, W. K. Tiong et al., "Utilisation of website logo for phishing detection," Computers & Security, vol. 54, pp. 16-26, 2015

[11] K. M. kumar, K. Alekhya, "Detecting phishing websites using fuzzy logic," International Journal of Advanced Research in Computer Engineering Technology(IJARCET), vol. 5, no. 10, 2016.