WEB PHISHING DETECTION

TEAM LEADER:
KALLEDA MANOJ KUMAR

TEAM MEMBERS:
NAVEEN R
KOTA HARI SRI RAGHAVENDRA
RAYALA VIJAY SAGAR

GUIDE:
Mrs. RAMYA J

INTRODUCTION:

An increasingly prevalent security hazard nowadays may be a cyberattack. Utilizing social engineering platforms, the attackers often aim to realize private information. Phishing has significantly increased online with the event of the internet over the past several years. The annual financial damage to Internet users from phishing is within the billions. The goal of phishing, a social engineering assault, is to spot the victim's personal information and steal it. It often uses email, text messaging, website spoofing, etc. We are unable to scale back the frequency of phishing attacks because attackers use classic approaches such ascending the target some alluring communications in order to deceive them and carry out an attack.

Phishing may be a common method attackers employ to send their payloads via emails. the bulk of users cannot figure out which received emails have a malicious attach mentor payload. during this case, Al and ML could play a pivotal role in identifying and preventing such phishing attacks. they will scan and identify phishing emails much faster than a human being can. they will also quickly differentiate malicious websites from legitimate websites. The researcher is thus concentrating on switching from the normal method to Artificial Intelligence approaches to safeguard internet access.

OBJECTIVE:

Phishing may be a social engineering assault that seeks to take advantage of flaws in system operations that have been caused by users of the system. A system could, as an example, be technically secure enough to stop password theft, but uninformed end If a hacker solicited people for his or her passwords, they could divulge them. Their passwords via a specified Protocol for Hypertext Transfer (HTTP) link, which eventually jeopardises the security of the programme. Additionally, technical flaws (such as name System By using (DNS)

cache poisoning), attackers Create plenty more persuasive social engineering messages Use of genuine but fake domain names can be extremely harmful (i.e., Influencing than using many domain names).

Due to this, phishing assaults are a multi-layered problem, and both technological and human-level issues must be addressed for an efficient mitigation to take place. Phishing attacks are challenging to resist because they aim human weaknesses (i.e., system end users). as an example, consistent with an evaluation in, even after receiving training from the top-performing user awareness programme, end users were unable to spot 29% of phishing assaults. The performance of software phishing detection approaches, on the opposite hand, is usually understood when it comes to targeted phishing attempts because they are tested against bulk phishing attacks. thanks to these shortcomings in phishing mitigation techniques, numerous firms, including top information security providers, have practically experienced security breaches.

LITERATURE SURVEY:

This chapter reviews the research on phishing attack detection. Phishing attacks target holes in systems that exist because of the involvement of humans. Users are the weakest link in the security chain since a large majority of cyberattacks are disseminated using techniques that take use of vulnerabilities detected in end users. Since there is no single, effective way to address all of the weaknesses in phishing, numerous strategies are frequently used to counteract particular attacks. Many of the recently proposed phishing detection strategies are discussed in this chapter.

R. Kiruthiga & D. Akila (2019) [1] in their research presented a novel method for detecting phishing websites using machine learning techniques. The accuracy of five machine learning techniques was also compared: Generalized Linear Model (GLM), Generalized Additive Model (GAM), Gradient Boosting (GBM), Random Forest (RF), and Decision Tree (DT). Accuracy, Precision, and Recall evaluation measures were computed for each method and contrasted. In contrast to R, which is an open-source computer language for performance analysis, Python is used to extract website information (30). A table was used to compare the effectiveness of the top three algorithms, Decision Tree, Random Forest, and GBM. According to the tables of accuracy, recall, and performance, the Random Forest algorithm had the highest 98.4% accuracy, 98.59% recall, and 97.70% precision.

Sumathi K & Sujatha V (2019) [2] found in their investigation that J48, Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB), and Artificial Neural Network (ANN) were among the machine learning techniques that were frequently used to identify phishing assaults. But one of the major challenges in machine learning is obtaining high-quality training data. In order to identify phishing Uniform Resource Locators, a deep learning technique known as Deep Neural Network (DNN) is introduced (URLs). A 30-dimension feature vector is first built using a feature extractor and is based on URL-based features, HTML-based features, and domain-based features. For the purpose of detecting phishing attacks, these features are input to the DNN classifier. It

has a single input layer, several hidden layers, and a single output layer. In DNN, there are numerous hidden layers that work to gradually learn high-level features. The DNN then produces a probability number that represents both valid URLs and phishing URLs. Using DNN increases phishing attack detection's precision, recall, and accuracy

Wang *et al.* [5] discovered that the phishing URL is treated as a string and that the text categorization challenge is identical to the problem of detecting phishing websites. In their suggested approach, we use machine learning to identify phishing websites and treat the problem of phishing website detection as a classification problem. With the deep learning approach, the neural network may extract the intrinsic feature expression in the URL data during the model-training phase and subsequently categorize the website as a phishing website. They make use of PDRCNN. A URL string is the input to PDRCNN, and its output is if the URL leads to a phishing website. When PDRCNN receives a URL string, it first encodes it as a string into a two-dimensional tensor of a fixed space before sending the encoded tensor to the deep learning neural network that was specifically constructed. The model first extracts the structural and semantic properties from the URL, classifies them using the Sigmoid function, and then outputs the classification of the URL.

Adebowale *et al.* [7] concentrated on the design and implementation of a deep learning-based phishing detection solution that made use of the URL shortener and website content including photos, text, and frames. In this study, a hybrid classification model known as the intelligent phishing detection system was created by combining the convolutional neural network (CNN) and the long short-term memory (LSTM) algorithm (IPDS). The CNN and LSTM classifiers were trained using 1 million URNs and more than 10,000 pictures to create the suggested model. Then, by taking into account a number of variables, including the type of feature, the quantity of misclassifications, and split issues, the sensitivity of the suggested model was calculated.

Aljofey *et al.* [3] In their paper gave, a quick deep learning-based solution model for phishing detection based on the URL of the website is given. This model uses character-level convolutional neural networks (CNN). The suggested model does not call for the use of any third-party services or the retrieval of any content from the target website. Without having any prior knowledge of phishing, it collects data and sequential patterns of URL strings, using the sequential pattern features for quick classification of the actual URL. Comparisons between numerous classical machine learning models and deep learning models are offered for evaluations utilizing a range of feature sets, including hand-crafted, character embedding, character level TF-IDF, and character level count vectors features. Results of the experiments indicate The suggested model outperformed the current phishing URL models, achieving accuracy of 95.02% on their dataset and 98.58%, 95.46%, and 95.22% on benchmark datasets.

From the literature survey ,we propose the usage of a hybrid algorithm with CNN and short term memory networks to build an AI model in web phishing detection.The proposed algorithm finds to be effective compared to the other

machine learning algorithms with the use of URL in detecting web phishing and based on this data an highly adaptive AI model can be build for the web phishing in an unorthodox method rather than our decade old classic method.

REFERENCE:

[1] Kiruthiga R, Akila D (2019) Phishing websites detection using machine learning. Int J Recent Technol Eng 8(2):111–114

[2] Sumathi K, Sujatha V (2019) Deep learning based-phishing attack detection. Int J Recent Technol Eng (IJRTE) 8(3):8428–8432

[3] Aljofey A, Jiang Q, Qu Q, Huang M, Niyigena JP (2020) An effective phishing detection model based on character level convolutional neural network from URL. Electronics 9(9):1514

[4] Mahdavifar S, Ghorbani AA (2019) Application of deep learning to cybersecurity: a survey. Neurocomputing 347:149–176

[5] Wang W, Zhang F, Luo X, Zhang S (2019b) PDRCNN: precise phishing detection with recurrent convolutional neural networks. Secur Commun Netw 2019:2595794.

[6] Basit A, Zafar M, Liu X, Javed AR, Jalil Z, Kifayat K (2020) A comprehensive survey of AI-enabled phishing attacks detection techniques. Telecommun Syst 76:139–154

[7] Berman DS, Buczak AL, Chavis JS, Corbett CL (2019) A survey of deep learning methods for cyber security. Information 10(4):122

[8] Adebowale MA, Lwin KT, Hossain MA (2020) Intelligent phishing detection scheme using deep learning algorithms. J Enterp Inf Manag. https://doi.org/10.1108/JEIM-01-2020-0036

[9] Hannousse A, Yahiouche S (2021) Towards benchmark datasets for machine learning based website phishing detection: an experimental study. Eng Appl Artif Intell 104:104347

[10] El Aassal A, Baki S, Das A, Verma RM (2020) An in-depth benchmarking and evaluation of phishing detection research for security needs. IEEE Access 8:22170–22192

[11] Feng J, Zou L, Nan T (2019) A phishing webpage detection method based on stacked autoencoder and correlation coefficients. J Comput Inf Technol 27(2):41–54

[12] Ferreira M (2019) Malicious URL detection using machine learning algorithms. In: Proc. Digit. Privacy Security Conf., pp 114–122

[13] Gangavarapu T, Jaidhar CD, Chanduka B (2020) Applicability of machine learning in spam and phishing email filtering: review and approaches. Artif Intell Rev 53:5019–5081

[14] Zamir A, Khan HU, Iqbal T, Yousaf N, Aslam F, Anjum A, Hamdani M (2020) Phishing web site detection using diverse machine learning algorithms. Electron Libr 38(1):65–80