WWW.IJRPR.COM

International Journal of Research Publication and Reviews

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

Detection of Phishing Website Using Machine Learning

S. Mahalakshmi[1], P. Meena[2], Mr. R. Gopinath[3]

[1]UG Student, Dept. of CSBS., E.G.S Pillay Engineering College, Nagapattinam, Tamil Nadu, India
[2]UG Student, Dept. of CSBS., E.G.S Pillay Engineering College, Nagapattinam, Tamil Nadu, India
[3]Assistant Professor, Dept. of CSBS., E.G.S Pillay Engineering College, Nagapattinam, Tamil Nadu, India
DOI: https://doi.org/10.55248/gengpi.5.0224.0519

ABSTRACT:

Phishing is a kind of cybercrime where fake emails, texts, and websites are used to steal credit card numbers, passwords, and other private information. With the growth of the internet and e-commerce, phishing attacks have gotten increasingly sophisticated and difficult for consumers to identify and avoid. One of the most efficient and widely utilized techniques employed by cybercriminals to steal our personal and financial information and con us out of money is phishing. In the present era, phishing assaults are more complex and difficult to detect. 97% of security specialists in an Intel survey claimed they were unable to distinguish between legitimate and phishing emails. Machine learning is an effective approach for detecting phishing websites. A large dataset of legitimate and fraudulent websites can be used to train machine learning algorithms, giving them the ability to distinguish between the two. This could help develop effective phishing detection systems that automatically identify and warn users of potentially dangerous websites.

KEYWORDS: Phishing, URL, Machine learning.

I. INTRODUCTION

These days, security specialists are particularly concerned about phishing since it is so easy to develop a phony website that closely resembles a real one.

Although experts can recognize phony websites, some people are unable to do so, and as a result, they fall prey to phishing attacks. The attacker's primary goal is to obtain bank account passwords. Businesses in the US lose $2 billion annually as a result of their clientele falling victim to phishing. According to the 3rd Microsoft Computing Safer Index Report, which was published in February 2014, the yearly global impact of phishing may reach $5 billion. Because users are growing less informed, phishing assaults are becoming more successful. It is highly challenging to combat phishing attacks since they prey on user vulnerabilities, yet improving phishing detection techniques is crucial.

The "blacklist" approach, which is another name for the basic technique of identifying phishing websites, involves adding IP addresses to the antivirus database and updating banned URLs. Attackers utilize inventive methods to trick people into thinking the URL is authentic by obfuscation and other straightforward tactics like fast-flux, which generates proxies automatically to host the webpage, algorithmic URL creation, etc., in order to avoid being included in blacklists. This method's primary flaw is that it is unable to identify phishing attacks that occur at zero hour. Heuristic-based detection, which is capable of identifying traits present in real-world phishing attempts, zero-hour phishing attack, yet there is a large false positive rate in detection and the features are not always present in such assaults.

Machine learning techniques have become the focus of many security researchers as a way to overcome the limitations of heuristics-based approaches and blacklists. Machine learning technology is made up of numerous algorithms that use historical data to forecast or make decisions about future data. With this method, phishing websites, including zero-hour phishing websites, may be accurately detected by the algorithm by analyzing the attributes of different blacklisted and legal URLs.

Web fraud comes in many forms, and phishing websites are frequently used as a point of entry for attempts at online social engineering. The hacker first constructs a webpage by pretending to be a trustworthy website. They then employ spam chats, texts, or social networking platforms to transmit these dubious URLs to prospective victims in the hopes that gullible people would take them for real [10]. Users' personal information (bank account numbers, government savings numbers, and so on) will be compromised if they submit it at the link that the hacker sends.

Numerous tactics exist to counteract phishing. Nearly every business has been greatly impacted by artificial intelligence (AI), including cyber security. This is because AI can identify attacks such as spear phishing, phishing, and spam historical attacks in the shape of databases.

In order to identify phishing domains, this study constructs and compares machine learning (ML) classification models. By utilizing the most accurate model among the four to determine whether a website is legitimate or a phish, detection is intended to be improved. Since phished domains entail social and technical challenges for which there is no universally applicable analysis, they are challenging to examine and understand. Therefore, in order to

better determine where to focus the model to lessen the danger emerging from visiting a phishing website—particularly with regard to consumer trust—all phishing domain causes and features were studied statistically and qualitatively.

II. RELATED WORK

Currently, there are three primary techniques of detecting phishing web pages: URL feature detection, webpage content feature detection, and blacklist-based detection. The blacklist detection feature is limited to collecting samples of phishing websites and updating the blacklist database on a timely basis. It only executes basic database query operations, and its detection rate is quick, easy, and convenient. In order to assess the legitimacy of a web page, content-based detection methods must first gather web material.

This is done by comparing the content of the page to previous online content or by using machine learning technology. The requirement to collect web content for this strategy raises the client's risk. It also necessitates a significant amount of manual feature engineering. Numerous elements require confirmation from pertinent experts. The caliber of the manually extracted features has a significant impact on its performance. Phishing attackers can effortlessly evade the detection model because of comparatively stable attributes. Current approaches mostly use neural networks to automatically extract URL features in order to verify the validity of web pages based on URL feature detection. In conclusion, two-word segmentation techniques and a single neural network are the primary methods used in the process of identifying phishing web pages based on URL attributes.

The following are this method's limitations: Word-based URL segmentation can result in a high number of words, which will increase the data set's features proportionately. Additionally, testing may not be able to collect the embedding vectors of newly arriving words; Partitioning the URL according to characters will result in certain delicate terms, such "login," "password," "registered," and so on, losing part of their useful information; A single neural network, like CNN, can only extract the local aspects of a URL; it is unable to extract the URL's sequence features. We suggest a technique based on sensitive word segmentation to address the issues with the previous approaches. This can clearly mark the important information in the URL, which helps the neural network classifier to extract more representative features.

The URL is divided into word levels according to the special characters, and the special characters are treated as words to obtain the effective information of the special characters. Next, the non-sensitive words are divided into character levels, and the sensitive words are regarded as a whole with the rest of the characters to distinguish. Convolutional neural networks and bidirectional long short-term memory networks are used by neural network classifiers to extract more abundant information from URLs.

III. PROPOSED SYSTEM

In today’s world, it has become increasingly difficult to distinguish between phishing emails and messages, even with the abundance of methods available today. Blacklisting, white listing, heuristics, and machine learning are a few of the methods used today to recognize phishing emails. This chapter offers a machine learning method to identify phishing emails and stop consumers from giving up their passwords, user IDs, or pins. In this study, we introduced several machine-learning techniques to identify phishing attempts in any format. Since the input URL may be categorized as either (1) phishing or (0) legal, we used the data set that is relevant to the classification problem. The following supervised machine learning models (classification) are being considered for this project in order to train the dataset:

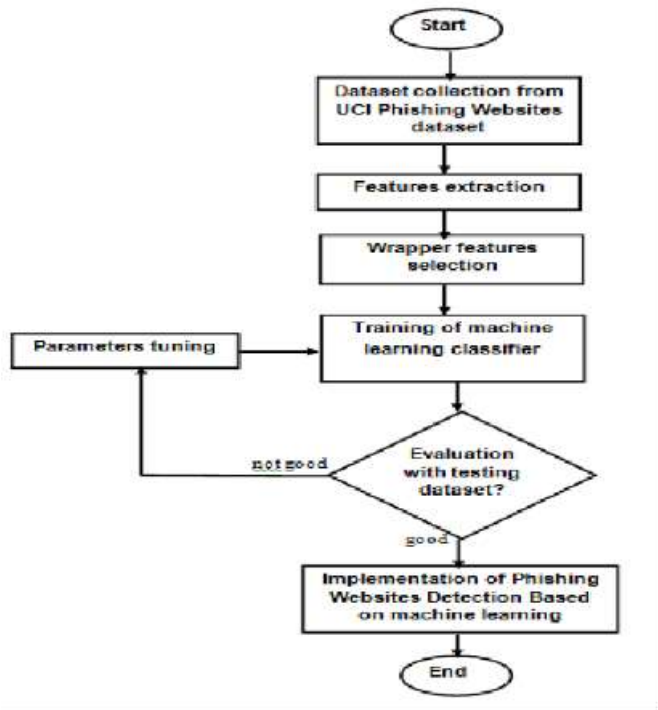

Fig 1: Flowchart of Proposed Solution

IV. MACHINE LEARNING MODELS

Machine learning challenge is supervised. Supervised machine learning tasks fall into two main categories: regression and classification. The input URL for this data set is categorized as either valid (1) or phishing (-1), indicating a classification error. For the purpose of training the data, the following machine learning models have been chosen: logistic regression, K-Nearest Neighbors, Support Vector Machine, and Gradient Boosting Classifier

4.1 Logistic Regression

A mathematical method called logistic regression is used to predict the likelihood of a binary response based on one or more independent variables. This indicates that a result with two values, like 0 or 1, pass or fail, yes or no, etc., can be predicted using logistic regression if specific factors are met. The logistic regression is a predictive analysis, just as the other regression models. It is commonly used to display data and highlight the relationship between one or more nominal, ordinal, interval, or ratio-level independent variables and one or more dependent binary variables. It also requires a more intricate cost structure. Instead of being a linear function, this cost function is referred to as the "sigmoid function" or "logistic function." The premise behind this algorithm

4.2 K Nearest Neighbors

The simplest algorithm is KNN, or K-Nearest Neighbors. It is capable of handling regression- and classification-based issues. It is widely utilized in decision-making systems, straightforward advisory systems, and image recognition technologies. KNN is used by online retailers such as Amazon and Netflix to recommend a range of publications to customers looking to purchase goods or rent movies. KNN operates on the foundation of accepted mathematical ideas. The conversion of data items into their precise values is the first step in the KNN implementation process. It functions in this way by highlighting the distance between these places' numerical rates. Subsequently, it calculates the likelihood of points being exactly the same as the testing data and groups them according to which points have the highest probability overall. When classifying data with KNN, the best frequency class among the K most similar occurrences is used to calculate the final result. For each new data instance, the standardized prevalence of instances that make up each class in the set of K closest similar cases is used to calculate the class probabilities

4.3 Support vector machine

Another potent algorithm in machine learning technology is the support vector machine. Each data item is plotted as a point in n-dimensional space by the support vector machine algorithm, which then creates a separation line to classify the data into two groups. They're known as Maximum Margin Classifiers because of their unique capability to contemporaneously minimize the empirical bracket error and maximize the geometric periphery. SVMs operate by mapping the input vectors into an advanced dimensional space where a minimal separation hyperplane is formed. They're grounded on the

Structural threat Minimization (SRM) principle. To divide the data, two resembling hyperplanes are erected on either side of the separating hyperplane. The separating hyperplane is determined by opting for the hyperplane that optimizes the distance between these resembling hyperplanes. It's believed that a lesser periphery or separation between the resembling hyperplanes improves the classifier's conception error

4.4 Gradient Boosting classifier

One machine learning algorithm that is a part of the ensemble methods class is the gradient boosting classifier. It is applied to classification challenges and creates a powerful classifier by merging several weak classifiers. The way this algorithm operates is by repeatedly adding decision trees to the model, each of which aims to improve upon the mistakes made by the preceding tree. Using the data, the algorithm first trains a weak classifier. Next, it determines the weak classifier's residuals—the difference between the values that were predicted and those that really occurred—and fits a new weak classifier to the residuals. After a predetermined number of iterations or until the residuals are sufficiently reduced, the process is repeated.

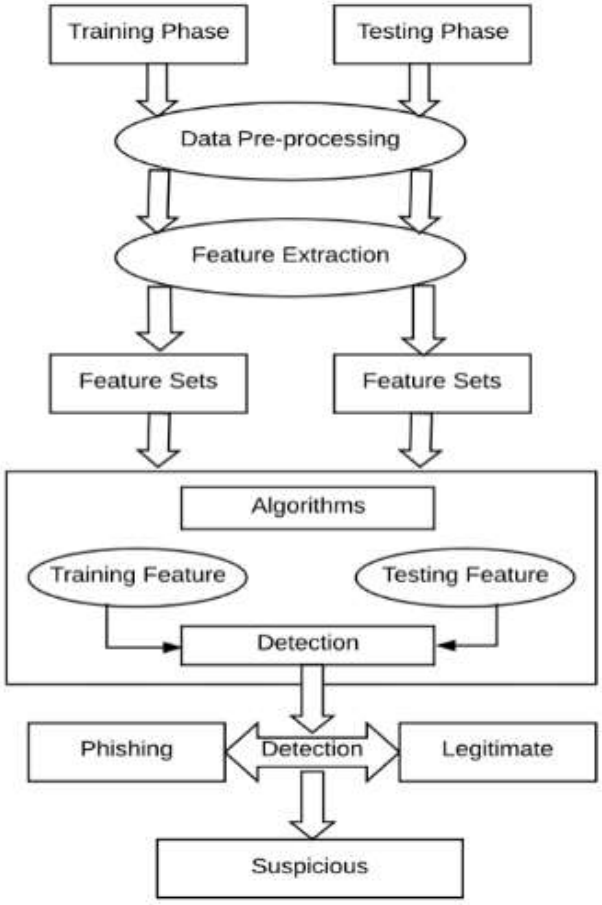

Fig 2: Phishing Architecture

V. FUTURE ENHANCEMENTS

Further developments using Gradient Boosting Classifiers (GBC) have great potential for phishing website identification. The improvement of feature engineering techniques represents one path forward. The discriminatory power of the model can be enhanced by continuously examining and expanding the feature set that was used for training it. This feature set can include more complex attributes obtained from website content analysis, like HTML structure and JavaScript code examination, in addition to more traditional indicators like URL length and HTTPS presence. Additionally, adding ensemble methods may result in even greater performance improvements. By utilizing a variety of model designs, combining Gradient Boosting with other ensemble techniques such as Random Forests, AdaBoost, or XGBoost may improve prediction accuracy. Furthermore, thorough hyperparameter tuning that is accomplished using techniques like grid search or randomized search can Optimize model performance by fine-tuning parameters such as learning rate and number of estimators. These developments should increase the effectiveness of phishing website detection systems, enabling them to counteract ever-more-advanced phishing techniques and better protect people from online dangers.

VI. CONCLUSION

Phishing attacks continue to be one of the main threats that individuals and companies must deal with nowadays. Phishing has increased in frequency on social media in tandem with its growth. Several supervised machine-learning algorithms were employed in this effort to identify phishing assaults. We gathered and categorized a collection of phishing URLs from a phish Tank. Consequently, the websites are represented by the classification model as either authentic or phishing. Figures 3 and 4 demonstrate that the URL is phishing and that the website is authentic, respectively. 94.2% is the greatest model performance achieved by the XGBoost Classifier, according to the results of the above mode. This project could be enhanced further by creating a graphical user interface (GUI) that analyzes web browsers. A URL to ascertain if it is legitimate or fraudulent. Creating powerful anti-phishing strategies that protect users

VIII. REFERENCES

[1] Chen, Sen, et al. 2019 GUI-squatting attack: Automated generation of Android phishing apps. IEEE Transactions on Dependable and Secure Computing.

[2] Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013

[3] Niu, Xiaofei, Guangchi Liu, and Qing Yang 2020 OpinionRank: Trustworthy Website Detection using Three Valued Subjective Logic. IEEE Transactions on Big Data

[4] Hassan Y.A. and Abdelfettah B, “Using case- based reasoning for phishing detection", Procedia Computer Science, vol. 109, 2017, pp. 281–288.

[5] Rao RS, Pais AR. Jail-Phish: An improved search engine-based phishing detection system. Computers & Security. 2019. Jun 1;83:246–67.

[6] Almseidin M, Zuraiq AA, Al-kasassbeh M, Alnidami N, “Phishing detection based on machine learning and feature selection methods”, International journal of interactive mobile technology, vol. 13, no. 12, pp. 171–183, 2019.

[7] Sahingoz OK, Buber E, Demir O, Diri B, “Machine learning based phishing detection from URLs”, Expert System Application, vol. 117, pp. 345–357, 2019.

[8] Zhu, E.; Ju, Y.; Chen, Z.; Liu, F.; Fang, X. DTOF-ANN: An Artificial Neural Network Phishing Detection Model Based on Decision Tree and Optimal Features. Appl. Soft Comput. 2020

[9] Tan CL, Chiew KL, Wong K, “PhishWHO: phishing webpage detection via identity keywords extraction and target domain name finder”, Decision Support Systems, vol. 88, pp 18–27, 2016

[10] Chiew KL, Chang EH, Tiong WK,”Utilisation of website logo for phishing detection”, Computer Security, pp.16–26, 2015.

[11] Zamir A, Khan HU, Iqbal T, Yousaf N, Aslam F et al., “Phishing web site detection using diverse machine learning algorithms”, The Electronic Library, vol.38, no.1, pp. 65–80, 2019

[12] Hong J., Kim T., Liu J., Park N., Kim SW, “Phishing URL Detection with Lexical Features and Blacklisted Domains”, Autonomous Secure Cyber Systems. Springer, 10.1007/978-3-030-33432-1_12.

[13] Gandotra E., Gupta D, “An Efficient Approach for Phishing Detection using Machine Learning”, Algorithms for Intelligent Systems, Springer, Singapore, 2021