IJIREEICE

ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁘ Peer-reviewed & Refereed journal ⁘ Vol. 12, Issue 5, May 2024

DOI: 10.17148/IJIREEICE.2024.12528

Phishing Website Classification Using Machine Learning

Kavita Nale[1], Swati Kulal[2], Ankita Masal[3], B.H. Patil[4]

Department of Electronics and Telecommunication, VPKBIET, Baramati

Abstract: A standard method of deceiving unwary individuals to disclose certain information is phishing. Personal information such as your username, password, and details of online financial transactions shall be collected through Phishing websites. The fishermen use their utilize websites with the identical appearance and language design as official websites. We must employ ant phishing techniques in order to identify phishing attempts as technology for their protection develops. This article discusses about detection strategies and approaches for detecting through machine learning. Attackers often use phishing because it's less difficult to trick victims into clicking harmful links which seem legitimate rather than trying to circumvent computer security.

Keywords: Phishing technique, Machine Learning, anti-phishing, phishing attacks, security and privacy, website features, categories, phishing approaches.

I. INTRODUCTION

Due to the big issue of developing a fake website that closely resembles that looks just like real website, phishing is becoming a major problem for security researchers. Even though professional are able to spot phony websites, some users are unable to do so and which is why some people fall victim of them, phishing scams. The primary objective of the attacker is to obtain login information for bank accounts. As users are unaware of phishing attacks, they are more and more successful. Phishing attacks are exceedingly hard to fight because they take advantage user vulnerabilities, yet it is improving phishing detection methods is critical. Phishing is a popular kind of extortion in which a malicious website impersonates a reliable company in order to get private data like passwords, login credentials, or MasterCard numbers. Despite this, phishers continue to using novel and hybrid approaches to get bypass open frameworks and programming, even though a number of techniques and tools that may identify possible phishing efforts in messages and typical phishing content on websites. Phishing is a type of fraud that involves the use of social engineering techniques and the dissemination of private and sensitive data, such as passwords and public credit information, by pretending to be someone else. Fake communications are created that look authentic and purport to be from real sources, such as financial institutions, online retailers, etc., in order to trick people into visit phony websites through links on phishing websites.

I. LITERATURE SURVEY

1.Phishing Website Detection in Real Time Amirh Abdullah, Nural, Abdulghani Ali Ahmed By using web spoofing, users are tricked into interacting with phony websites instead of authentic ones. The primary goal of these attacks is to steal users' sensitive information. A "shadow" website that loops similarly to the official website is created by the attacker. The attacker can view and alter any data that the user provides them with thanks to this fraudulent act. This paper proposes a method for phishing website detection that involves looking at the URLs of web pages. Through a uniform resource locator (URL) check of the suspected web page, the proposed solution can differentiate between a legitimate and a fake website. URLs are examined according to specific criteria in order to verify if they are phishing websites. Any attacks that are found are reported in order to stop them the demonstration.

2.Detection of Phishing Attacks Muhammet Baykara, Zahit Ziya Gurel Phishing is a type of cybercrime in which a perpetrator poses as a legitimate organization or person and uses email or other communication channels to promote themselves as such. In this type of cyberattack, the perpetrator uses phishing emails to send harmful links or attachments that can carry out a variety of tasks, such as obtaining the victim's account information or login credentials. The victims of these emails suffer from identity theft and financial loss. A program known as the "Anti-phishing Simulator" was created for this study and provides information on phishing detection issues, including how to identify phishing emails. This software examines the contents of emails to detect phishing and spam emails. The Bayesian algorithm classification of spam words added to the database is given.

3. WC-PAD: Phishing Attack Detection Nathezhtha based on Web Crawling Sangeetha, T1.D2, Vaidehi.V3 Phishing is a forbidden activity in which private information about users is stolen. Phishing websites aim to harm individuals,

companies, government websites, and cloud storage hosting companies. Even though software-based techniques are favored due to operational and budgetary reasons, hardware based approaches are currently widely used for anti-phishing. the current attacks on phishing detection websites. A three-phase attack detection system known as the Web Crawler based Phishing Attack Detector (WC-PAD) has been proposed as a solution to these problems and a more accurate way to identify instances of phishing. It classifies websites as phishing or non-phishing based on input features such as web traffic, web content, and Uniform Resource Locator (URL). The WC-PAD is subjected to an experimental analysis using datasets gathered from actual phishing cases derived from the trial.

4. Detection of Phishing Attacks with Machine Learning Techniques in Cognitive Security Architecture Ivan Ortiz-Garces, Robert O. Andrade, and Mar'1a Cazares Phishing attacks have become more frequent in Latin America than cybersecurity analysts' operational capabilities. The application for cognitive security suggests using data analytics, machine learning, and big data to speed up attack detection response times. The analysis of anomalous behaviour associated with phishing web attacks is presented in this paper, along with a discussion of machine learning techniques as a potential solution. In order to minimize the impact of an attack, this analysis of URLs is based on specific characteristics of the URLs and aims to provide real-time information.

5. A survey of the QR code phishing: the current attacks and counter measures Kang Leng Chiew and Choon Lin Tan After gaining traction, quick response (QR) codes were modified for a number of uses, including authentication and as a pointer to digital data. Although the code serves as a convenient physical pointer to the online world, it can be tricked into directing a link's intended destination to a malicious website. Thus, phishers can easily launch phishing attacks by utilizing QR codes. Here, phishing attempts that currently use QR codes as a vector are also reviewed. Additionally, it is discovered that the current countermeasures are inadequate and vulnerable to issues like limited data space in the code, high overhead solutions, and bar-code-in-barcode attacks. When compared to the amount of effort put into email and web phishing detection, QR code phishing detection is less extensive.

II. ALGORITHM

Random Forest:

Random forest techniques are among the machine learning algorithms. While random forest techniques can be utilized to regression and classification problems as well, they are most commonly used to classification problems. This classifier utilizes the output of a decision tree that it builds. The largest tree that yields the same result determines the output. One popular machine learning algorithm is Random Forest, which used in supervised learning techniques. It can be used to ML problems solving both classification and regression. Its basis the concept of ensemble learning, which is the idea of process of combining multiple classifiers to solve a difficult problem and improve the model's functionality.

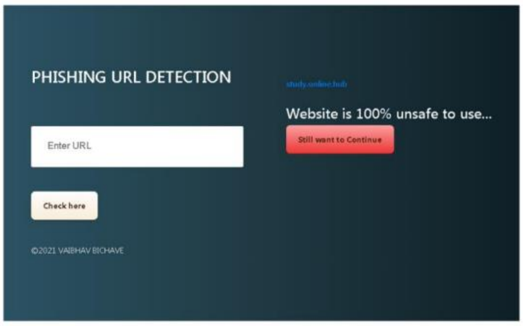
As implied by the name, “Random Forest” is a classifier that builds a number of decision trees on different subsets of the given dataset and takes the average to increase the predictive accuracy of that dataset. Rather than depending on a single decision tree, the random forest predict the outcome based on the majority vote of projections from each tree.

III. WORKING

Our first step is to take an unstructured dataset from places like GitHub, Kaggle, etc. One of the most important steps in the modelling process is data preprocessing because we are working with unstructured data. As a result, we extract certain data from the URL, including the domain name, length, and age of the URL. Since machine learning can only process numerical data, we extract the properties from the URL and give them a value, like 0 or 1. The model is trained using the decision tree and random forest approaches in the next phase, and their respective performances are compared. It is evident that applying both techniques results in increased accuracy when using random forests. and accuracy in contrast to decision trees.

However, a small pop-up alerting the user to this dangerous website will show on their screen if they enter a URL that turns out to be a phishing website. You can select "CONFIRM" to gain access to the website if a user needs to contact data on that page; otherwise, you will be redirected to the previous page.

DOI: 10.17148/IJIREEICE.2024.12528



PHISHING URL DETECTION

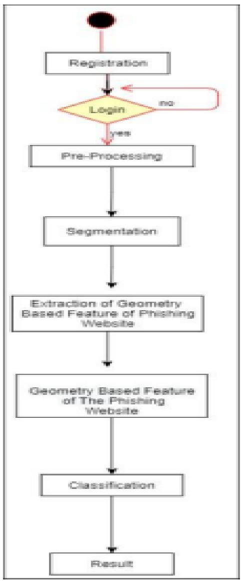
study.online.hub

Website is 100% unsafe to use...

©2021 VAIBHAV BICHAVE

Fig. URL of Website Checking Portal

IV. FLOWCHART



Registration
Login
no
yes
Pre-Processing
Segmentation
Extraction of Geometry Based Feature of Phishing Website
Geometry Based Feature of The Phishing Website
Classification
Result

V. RESULT

The SCIKIT-learn tool has been used to import machine learning algorithms. A testing set is used to gauge the classifiers' performance after each one has been trained using a training set. Each classifier's accuracy score has been utilized to assess how well it performed.

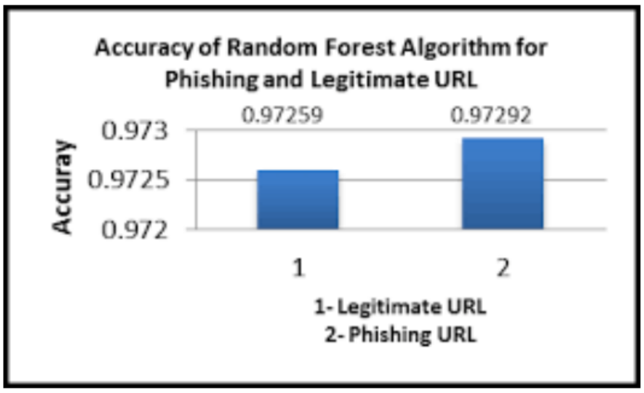


Fig. Accuracy of Random Forest

DOI: 10.17148/IJIREEICE.2024.12528

REFERENCES

[1]. In 2019, Natheztha.T, Sangeetha.D, and AIDEHI.V published"WC-PAD: Web Crawling based Phishing Attack Dection

[2]. "Detection of Phishing Attacks using Machine Learning Techniques in Cognitive Security Architecture," Ivan Ortiz-Graces, Roberto O. Andrade, and Maria Cazares, 2019.

[3]. "A summary of QR code phishing, including recent attacks and countermeasures" Kelvin S.C. Yoong, Choon Lin Tan, and Kang Leng Chiew. In 2019.

[4]. "A Cryptographically Secure AntiPhishing Tool for QR Code Attacks: Quick Response Code Secure," by V. Mavroeidis and M. Nicho, presented at the International Conference on Mathematical Methods, Models, and Architectures for Computer Network Security. 2017 Springer.

[5]. "Phishing web sites features classification based on extreme learning machine," 6th Int. Symp. Digit. Forensic Secur. ISDFS 2018 - Proceeding, vol. 2018–Janua, 2018, Y. So¨ nmez, T. Tuncer, H. Go¨ kal, and E. Avci.

[6]. "On Feature Selection for the Prediction of Phishing Websites," IEEE 15th International Conference on Dependable, Auton. Secur., and Computing, 15th International Conference on Pervasive Intelligence and Computing, 3rd International Conference on Big Data Intelligence and Computing, Cyber Sci. Technol. Congr., 2017.