ISSN No:-2456-2165

Enhanced Phishing Website Detection: Leveraging Random Forest and XGBoost Algorithms with Hybrid Features

Prof. Ashwini Bhavsar[1], Adarsh Waikar[2], Ayush Petkar[3], Seema Mane[4], Vishwatej Sarwale[5]
[1]Prof. Ashwini Bhavsar, Dept. of Computer Engineering, PCCOER, Maharashtra, India
[2]Adarsh Waikar, Dept. of Computer Engineering, PCCOER, Maharashtra, India
[3]Ayush Petkar, Dept. of Computer Engineering, PCCOER, Maharashtra, India
[4]Seema Mane, Dept. of Computer Engineering, PCCOER, Maharashtra, India
[5]Vishwatej Sarwale, Dept. of Computer Engineering, PCCOER, Maharashtra, India

Abstract:- Phishing technique is used by hackers or attackers to scam the people on internet into giving private details such as login credentials of various profiles, social security numbers (SSNs), banking information, etc. Attackers disguise a webpage as an official legit website. Blacklist or whitelist, heuristic, and visual similarity-based anti-phishing solutions are unable to detect zero-hour phishing assaults or newly created websites. Older methods are more complex and not suitable for day-to-day scenarios since they rely on external sources such as search engines. As a result, finding newly constructed phishing websites in a real-time context is a significant hurdle in the field of cybersecurity. This paper presents a hybrid feature-based anti-phishing approach that nullifies these problems by extracting characteristics from URL and hyperlink data that is only available on the client side. Also, a brand-new dataset is created for experiments employing well-liked machine-learning classification techniques. Our experimental findings dictated that the presented random forest-based phishing website detection approach is more effective and gives a higher accuracy result of 96.81% with the blend of the XG Boost technique.

Keywords:- *Cybersecurity, Phishing Detection, Machine Learning, Hyperlink Feature, URL Feature, Anti-Phishing, XG Boost, Hybrid Feature.*

I. INTRODUCTION

In 2022 alone, about 69% of the world's population, actively used the internet. This shows that number of internet users will keep on increasing in the coming times. In the field of cybersecurity, phishing is currently one of the most serious and dangerous online threats [1]. The rapid advancement of Internet technology has greatly boosted the use of social media, online banking, e-commerce services, and other similar services. In 2022, 166,187,118 harmful email attachments were stopped by Kaspersky Mail Anti-Virus. Aims to click on phishing URLs were blocked 507,851,735 times by our anti-phishing system. The takeover of a Telegram account was related to 378,496 attempts to click on phishing URLs. According to "A Digital Report in 2021" data from We Are Social (Global Overview Report 2021) [2], there are 4.66 billion internet users worldwide, up 7.3 percent (316 million additional users) from January 2020. Internet penetration currently stands at 59.5 percent, which gives phishing attackers the chance to profit by extorting and stealing private data from online users [3]. The attacker creates a fake website and distributes links via emails, Facebook, Twitter, and other social media applications. When a user unknowingly opens the link and changes or fills in any sensitive and private credentials, attackers obtain access to the user's information such as financial information, personal information, login credentials, and so on. Cybercriminals utilize stolen information for a range of illicit actions, including blackmailing victims. Consumers fall prey to phishing mainly because of the following reasons:

- User's understanding of URLs is generally poor
- Visitors do not know which websites to believe.
- Redirected, shorten URLs or hidden URLs prevent users from seeing the full address of the web page.
- Users do not have much time to look up a URL fast or unintentionally reach certain online pages.
- Consumers lack the ability to discern the difference between trustworthy and counterfeit websites.

Phishing assaults are now being used to distribute dangerous software such as ransomware. So, in this work, we concentrate on efficiently identifying phishing websites to prevent unaware internet users from falling victim to phishers and thereby lessen the emotional and financial damages. As of today, everything in our day-to-day lives is now digitally stored as data and the formally actionable insights that can be extracted are the reason to provide intelligent solutions. "Data science" has recently become a trending topic in the computing world. Such data-driven solutions may be utilized to create an effective model as well as an intelligent decision-making system in a variety of real-world application domains, such as business, financial analysis, cybersecurity, IoT applications, and many more. As a result, the goal of this article is to provide an effective data-driven solution that uses machine learning techniques to evaluate whether a website is phishing. The majority of machine learning-based phishing detection algorithms gather characteristics from the URL, search engine, third-party, online traffic, DNS, and so on.

ISSN No:-2456-2165

Because of the difficulties and time limits, these methods may not be suited for real-time phishing detection. Phishing sites have a typical life cycle of less than nine hours, and half of them are removed in less than a day. Yet, most phishing pages that use hacked domains remain online for longer than a day. As a result, the research question addressed in this work is, "How can we design an efficient and intelligent phishing detection model while taking into consideration the challenges listed above?"

In this study, we develop a hybrid feature-based phishing detection method to address this research topic. This method successfully detects phishing websites and addresses the aforementioned issues. We employ URL-based features to identify fraudulent websites.

Our feature extraction technique is independent of any search engine or third-party services. From the website's source code, we examine and extract Hyperlink characteristics. We get distinct types of data from the hyperlink information and different attributes from the URL. To train our classification model, we mix all the characteristics to create a hybrid feature set of features. The main contributions of our paper are as follows:

- We start by gathering legal and phishing website URLs from open-source platforms to construct a dataset. Phish Tank is such a platform for phishing website information.
- We provide a method for accurately detecting phishing that dynamically harvests hybrid characteristics and makes extensive use of them.
- Our suggested machine learning-based solution accurately and efficiently recognizes zero-hour phishing assaults.

II. LITERATURE SURVEY

In 2022, Sumitra Das Guptta, Khandaker Tayef Shahriar & Hamed Alqahtani [1] proposed a machine learning model and system that helps in detecting phishing websites by utilizing and analyzing URLs and their hyperlink-based complex features to get higher accuracy without using third-party applications. The classification algorithms used in the research are Logistics Regression, Random Forest, Decision Tree, SVM, and XG Boost. An accuracy of 99.17% is achieved over traditional approaches.

In 2022, Adarsh Mandadi; Saikiran Boppana; Vishnu Ravella; R Kavitha [2] this research involves the use of mainly two classifiers which are Random Forest and Decision Tree classifier. It correctly identifies the phishing and legitimate URLs with 87.0% and 82.4% accuracy respectively.

In 2020, Jian Feng, Lianyang Zou, Ou Ye and Jingzhou Han [3] proposed a Multidimensional Algorithm where the major components are automated learning representations from multi-aspect features through the representation learning and feature extraction using a hybrid model of deep learning network. This approach makes use of CNN-LSTM and NLP.

In 2020, Poonam Kumari, Apoorva H R Gowda, Bhandhavya K, Bhavya M U, Spurthi M N [4] proposed a Hybrid Mode to address the problems brought on by phishing websites. By merging several models, a hybrid-based model is created, which increases the accuracy of phishing attack detection.

In 2021, Sinduja. S, Monisha. S, Priya Dharshini. S, Sneha. K, Vaishnavi. R [5] proposed a Machine Learning Algorithm for efficient detection of phishing website using the hyperlink features and random forest classifier.

In 2021, [6] Om Sapate, Sumit Kolhe, Shantanu Taro, Vishal Kumar Kashyap proposed the framework for phishing website detection in mobile devices which exploits various web browsers plug-in as well as machine learning based engine to detect zero-hour phishing website.

In 2020, Mehmet Korkmaz, Ozgur Koray Sahingoz and Banu Diri [7] proposed a machine learning-based phishing detection system that analyses URLs using eight distinct algorithms and three different datasets to compare the findings to existing research.

In 2019, Mohammad Mehdi Yadollahi, Farzaneh Shoeleh, Elham Serkani, Afsaneh Madani, Hossein Gharaee [8] proposed a machine learning algorithm that can distinguish between legal and fraudulent websites online and with lots of features. The suggested method is totally client-side and doesn't call for any third-party services because it extracts various kinds of discriminative information from URLs and webpage source code.

In 2021, Youness Mourtaji, Mohammed Bouhorma, Daniyal Alghazzawi, Ghadah Aldabbagh and Abdullah Alghamdi [9] proposed a new approach that merged the scores of different features selected using the various feature selection methods hence increasing the dependency of the selected feature sets.

In 2020, Xuqiao Yu [10] proposed a Hybrid Model combining the Deep Belief Network and the machine learning method of Support Vector Machines to increase the accuracy of the detection. The Hybrid Model also covers up for the weakness of the other model used in the research.

III. METHODOLOGIES

Existing Methods: This method involves finding the websites credibility by finding the URL structure. Illicit domain or a phishing website looks suspicious based on the various reasons like if it has some misspelled words or is pointing towards a false top-level domain, involvement of fraudulent URL, a young domain age, significantly lower page rank, or long URL.

List-based techniques involves the use of whitelisted and blacklisted websites stored on universal website database such as phish Tanks. If the domain of the concerned website matches one present in the blacklisted sites, than it is termed as a Phishing Website and if the domain matches one present

ISSN No:-2456-2165

in the whitelisted sites than it is identified as a legitimate website.

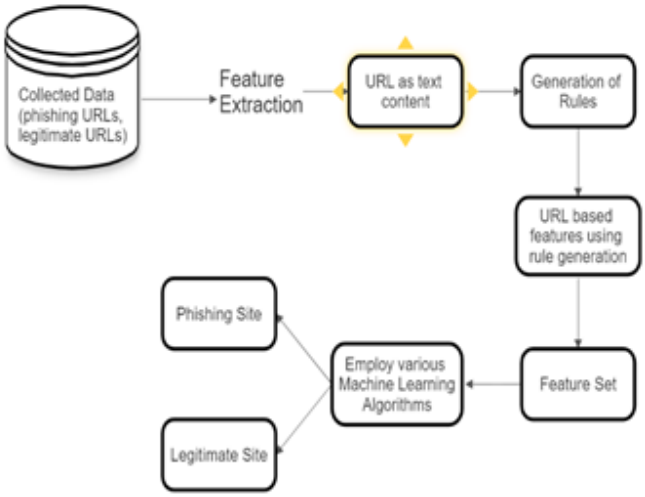

Fig 1 Existing Methodology

This method has a primary flaw since it cannot identify the newly created websites whose domains are absent from the listed database i.e.; it is not useful to deal with the zero-hour phishing attacks.

❖ *Proposed Method*

The List based characteristics, Visual Similarity based characteristics, Machine Learning based approaches help us to identify whether the website is valid URL or not. The various features category can be divided into four main categories:

A. Address Bar-Based Features

These features include those which are directly compiled from the URLs, like the URL length greater than 54, or whether an IP address is present in the URL, whether various URL shortening services (tinyurl.com or bit.ly) were used, or redirection is used. Additional features also include the following:

- Addition of suffix or prefix separated by (-) in the domain
- Presence of sub-domains and domain
- Existence of HTTPS
- Domain registration age
- Favicon loading from different Domain
- Using a non-standard port

B. Abnormal Features: These include

- Images are loaded in the body from a different URL
- Lesser or minimum use of meta tags
- Server Form Handler (SFH) uses
- Submitting information to the email
- An abnormal URL

C. HTML and JavaScript-Based Features: These includes the characteristics like:

- Website forwarding
- Page source code, photos, textual content used in the website
- HTML codes, CSS (Cascading Style Sheets), website logo, and so on
- Status bar customization, where a fake URL is displayed using various JavaScript methods
- Disabling the ability to right-click feature to prevent users from inspecting or viewing the source code
- Use of pop-up windows
- iFrame redirection

D. Domain-Based Features: These include:

- Doubtful young domains
- Suspicious DNS record
- Low website traffic
- PageRank, since it is observed that most of the phishing websites have no PageRank
- Site is indexed by Google or not

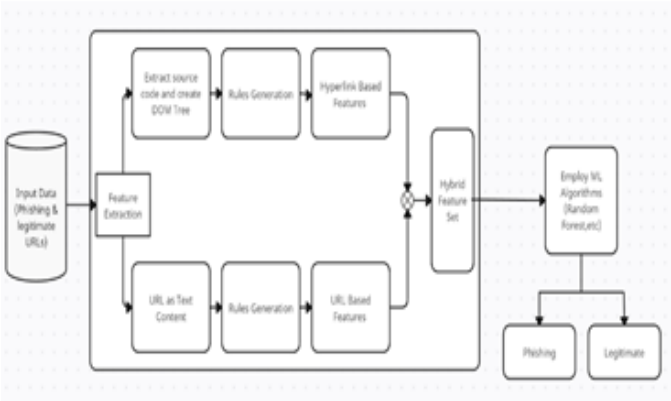

Fig 2 Proposed System

The above characteristics are based on the URL and hyperlink features of a website.

Building a machine learning model is the next step which helps us to detect the zero-hour phishing websites.

Given all the standards that can help us in detecting phishing URLs, we can use a machine learning algorithm, such as random forest classifier or a decision tree classifier to help us decide whether an URL is valid or not.

Machine Learning based Approach is used wherein a dataset is created with extracted features. Furthermore, a classification algorithm is trained on the URL and the Hyperlink characteristics of the phishing website. When a machine learning model is trained against heuristic features then it can also be used to detect the zero-hour phishing website. Overall, of all the phishing website detection approaches present, the machine learning approach is better suited.

❖ *Abbreviations and Acronyms*

A. CART

Defined as a Classification and Regression Tree (CART), is a special type of Decision Tree that describes how the values of a target variable can be predicted based on the values of feature variables.

B. URL

Defined as Uniform Resource Locator is an Internet resource that can be found using a URL (Uniform Resource Locator), a special identifier. It is additionally known as a web address.

C. SL

Defined as Supervised Learning is a machine learning technique used to train various models based on labeled data. Examples of Supervised learning algorithms are Logistics Regression, Linear Regression, and Naive Bias

D. RAM

Defined as Random-Access Memory is a type of computer memory where the data which is currently under process is stored or recorded. It stores the current working process data.

IV. CONCLUSIONS

From the study of various papers, we analyzed that out of the various methods present for the detection of phishing websites the machine learning method coupled with the URL and Hyperlink features gives the best result. Such a model is also capable of identifying zero-hour phishing websites which are often misinterpreted when analyzed by the existing methods.

REFERENCES

[1] K. T. S. Sumitra Das Guptta," Modeling Hybrid Feature-Based Phishing Websites Detection Using Machine Learning Techniques," Annals of Data Science, 2022.
[2] Odeh, Ammar Jamil et al. "Machine Learning Techniques for Detection of Website Phishing: A Review for Promises and Challenges." 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC) (2021)
[3] M. S. Sinduja. S, "Efficient Phishing Website Detection using Machine Learning Algorithm." International Journal for Research in Applied Science Engineering Technology, 2021
[4] Rishikesh Mahajan, "Phishing Website Detection using Machine Learning Algorithms", International Journal of Computer Applications Volume 181 – No. 23, October 2020
[5] Steve Sheng, Ponnurangam Kumaraguru, "Improving Phishing Countermeasures: An Analysis of Expert Interviews", International Journal of Computer Applications Volume 181 – No. 23, October 2020.
[6] Mohammad Alsharaiah, Ahmad Adel, "A new phishing website detection framework using Ensemble classification and clustering," International Journal of Data and Network Science, 2023
[7] Poonam Kumari, Apoorva H R Gowda, Bhandhavya K, Bhavya M U, Spurthi M N, "Detecting Phishing Sites using Hybrid Model," INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) 2020
[8] Om Sapate, Sumit Kolhe, Shantanu Taro, Vishal Kumar Kashyap, "Preventing Phishing Attacks in Real-Time Using Machine Learning," International Journal of Advanced Research in Science, Communication and Technology (IJARSCT) 2021.
[9] Jian Fend, Lianyang Zou, Ou Ye, Jingzhou Han, "Phishing Webpage Detection Method Based on Multidimensional Features Driven by Deep Learning," IEEE Access, 2020.
[10] Arathi Krishna V, Anusree A, Blessy Jose, Karthika Anilkumar, Ojus Thomas Lee, "Phishing Detection using Machine Learning based URL Analysis: A Survey," INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH TECHNOLOGY (IJERT) NCREIS – 2021