IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11695

DETECTION OF PHISHING WEBSITES USING MACHINE LEARNING

Anuja Bhosale[1], Gayatri Gadas[2], Muskan Chavan[3], Neha Pandhare[4], Seema Hadke[5]

[1,2,3,4]Student, IT Dept., Bharati Vidyapeeth's College of Engineering for Women, Pune, India

[5]Assistant Professor, IT Dept., Bharati Vidyapeeth's College of Engineering for Women, Pune, India

Abstract: Phishing websites that anticipate to take the victim's confidential data by diverting them to surf a fake website page that resembles a sincere to goodness one is some another type of criminal activity through the internet and its one of the especially concerns in numerous areas including e-managing an account and retailing. Detecting phishing sites is a complex and unpredictable process involving numerous variables and criteria that are not stable. Using Extreme Learning Machines, we proposed an intelligent model for detecting phishing web pages. There are different types of web pages with different features. Therefore, we must use a specific set of features on web pages to protect against phishing. A machine learning model was proposed to detect phishing web pages. To detect phishing web pages, we proposed a machine learning model. This study aims to detect phishing URLs and narrow down the best machine learning method based on accuracy, false-positive rate, and false-negative rate. Phishing, Feature Classification, Random Forest Classifier, and other terms are used in this study.

Keywords: Phishing attack, Machine Learning, Random Forest, Feature Classification, URL.

I. INTRODUCTION

Technology is advancing at a breakneck pace, and the internet has become an indispensable part of people's daily lives as a result. Because of the rapid advancement of technology and the widespread use of digital systems, internet use has increased, and data security has become increasingly important. The basic goal of information technology security is to ensure that adequate precautions are taken against threats and dangers that may be encountered by users when using these technologies.

Phishing is the deception of a trustworthy person in an electronic connection to get sensitive information such as usernames, passwords, and credit card numbers. It's usually done through email spoofing or instant messaging, and it often urges consumers to enter personal information on a bogus website that looks and feels exactly like the real one. Information security dangers have been observed and developed over time as the internet and information systems have evolved. The impact is a breach of information security due to the compromise of private data, with the victim potentially losing money or other assets as a result. Internet users are vulnerable to a variety of cyber risks, including the theft of personal information, identity theft, and financial losses. As a result, internet use in the home and at work may be questionable. To lessen security vulnerabilities, users must be able to identify and defend against privacy leakage using effective analytical tools. At the time of an attack, effective systems that can improve self-intervention must be built utilizing an artificial intelligence-based information security management system [17].

II. LITERATURE SURVEY

1. Amani Alswailem Bashayr Alabdullah Norah Alrumayh Dr. Aram Alsedrani, "Detecting Phishing Websites Using Machine Learning", IEEE 2019.

The system is based on a machine learning method, particularly supervised learning. Here is selected the Random Forest technique due to its good performance in classification. The focus is to pursue a higher-performance classifier by studying the features of phishing websites and choosing the better combination of them to train the classifier. As a result, the conclusion is the paper is with an accuracy of 98.8[2].

2. Alfredo Cuzzocrea, Fabio Martinelli, Francesco Mercaldo, "A Machine-Learning Framework for Supporting Intelligent Web-Phishing Detection and Analysis", ACM 2019.

In particular, the system makes use of state-of-the-art decision tree algorithms for detecting whether a Web site can perform phishing activities. If this is the case, the Web site is classified as a Web-phishing site. The experimental evaluation confirms the benefits of applying machine learning methods to the well-known web-phishing detection problem [3].

DOI: 10.17148/IJARCCE.2022.11695

3. Yasin S¨onmez1 T¨urker Tuncer2 H¨useyin G¨okal 3 Engin Avc4, "Phishing Web Sites Features Classification Based on Extreme Learning Machine.", IEEE 2018

The purpose of this study is to perform Extreme Learning Machine (ELM) based classification for 30 features including Phishing Websites Data in the UC Irvine Machine Learning Repository database. For results assessment, ELM was compared with other machine learning methods such as Support Vector Machine (SVM), and Naive Bayes (NB) and detected to have the highest accuracy of 95.34 percentage [6]

4. Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. Chaudhery, "Intelligent Phishing Website Detection using Random Forest Classifier", International Conference on Electrical and Computing Technologies and Applications (ICECTA), 2017

In this paper, an intelligent system to detect phishing attacks is presented. We used different data mining techniques to decide categories of websites: legitimate or phishing. Different classifiers were used to construct an accurate intelligent system for phishing website detection. Classification accuracy, the area under the receiver operating characteristic (ROC) curves (AUC), and F- measure is used to evaluate the performance of the data mining techniques.

Results showed that Random Forest has outperformed best among the classification methods by achieving the highest accuracy 97.36 percent. Random forest runtimes are quite fast, and it can deal with different websites for phishing detection [7].

5. Andrew J. Park Ruhi Naaz Quadari Herbert H. Tsang, "Phishing Website Detection Framework Through Web Scraping and Data Mining", IEEE 2017

The focus of this research is to establish a strong relationship between those identified heuristics(content-based) and the legitimacy of a website by analysing training sets of websites (both phishing and legitimate websites) and in the process analyse new patterns and report findings. Many existing phishing detection tools are often not very accurate as they depend mostly on the old database of previously identified phishing websites for this framework, a web crawler was developed to scrape the contents of phishing and legitimate websites. These contents were analysed to rate the heuristics and their contribution scale factor toward the illegitimacy of a website. The data set collected from Web Scraper was then analysed using a data mining tool to find patterns and report findings. A case study shows how this framework can be used to detect a phishing website. This research is still in progress but shows a new way of finding and using heuristics and the sum of their contributing weights to effectively and accurately detect phishing websites [8].

III. SYSTEM ARCHITECTURE

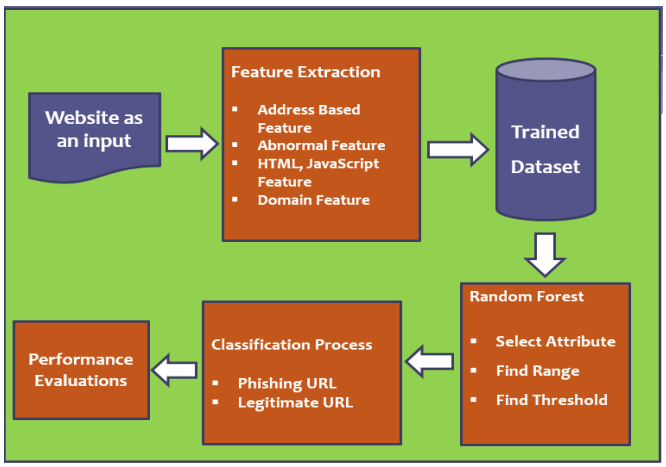

Fig. 1 System Architecture

The proposed methodology involves pre-processing imported data and importing a dataset of phishing and legal URLs. Four different types of URL features are used to detect phishing websites: domain-based, address-based, abnormal-based, and HTML and JavaScript features. With the processed data, certain URL features are extracted, and values for each URL attribute are generated. The URL is analysed using a machine learning algorithm that calculates the range value and threshold value for URL attributes. The URL is then classed as phishing or authentic. The attribute values are computed using phishing website feature extraction and are used to determine the range and threshold value. The values for each phishing attribute range from -1 to 1, with low, medium, and high values based on the phishing website feature. The classification of phishing and legitimate websites is based on the values of attributes extracted using four types of phishing categories and a machine learning approach [1][9].

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11695

IV. ALGORITHM

Random Forest:

Using the following steps, we can understand how the Random Forest algorithm works [1][7].

Step 1: First, start by selecting random samples from a dataset.

Step 2: Next, for each sample, this algorithm will construct a decision tree.

Each decision tree will then produce a prediction result.

Step 3: Each predicted result will be voted on in this step.

Step 4: Lastly, select the prediction result that has received the most votes.

V. IMPLEMENTATION

A dataset obtained from kaggle.com is being used where it is then split into training and testing datasets. Machine learning classifiers such as random forest and SVM come into play in this phase. After inserting the URL in the browser, we shall pass it through the browser is then broken down into different features and those are extracted. These features are compared with various stored patterns. After analyzing the extracted features, a decision of whether the input URL is phished or legitimate is made. The output is then presented to the user so that he/she can decide whether to leave the site or to stay.

Fig 2 shows the chrome extension created by using the training model marked in the red circle and you can see the URL of the webpage which you are searching.

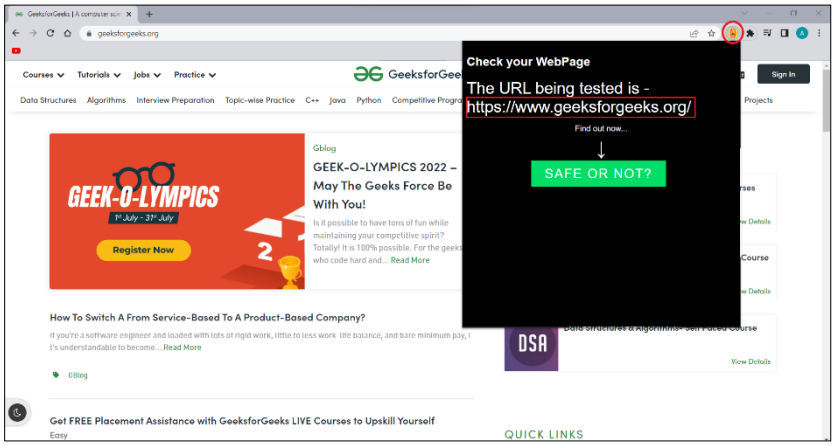

Fig. 2 Chrome Extension

Fig 3 below is an image of the output of a legitimate URL. After using the browser extension icon (circled in red), the pop window appears which lets the user know of the searched website is being phished or not, legitimate in this case (rectangle red marking)

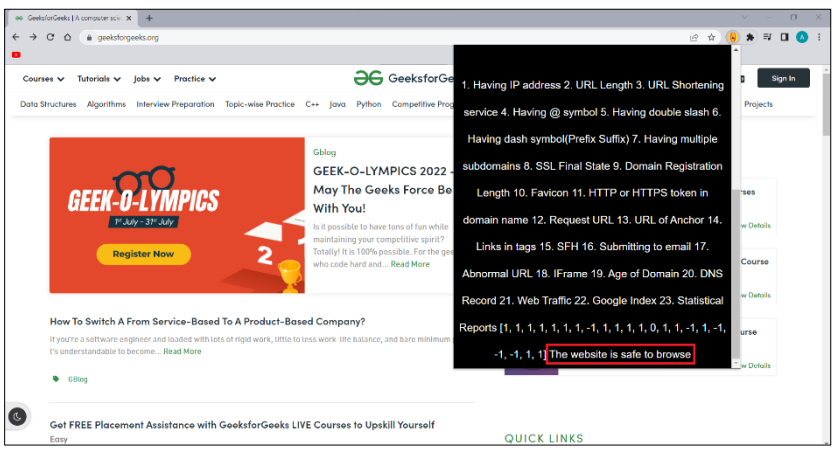

Fig. 3 Legitimate Website

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022

DOI: 10.17148/IJARCCE.2022.11695

On the other side, Fig 4 is a snapshot of the phishing URL and its output marked in the red rectangle which warns users to do not to visit this website.

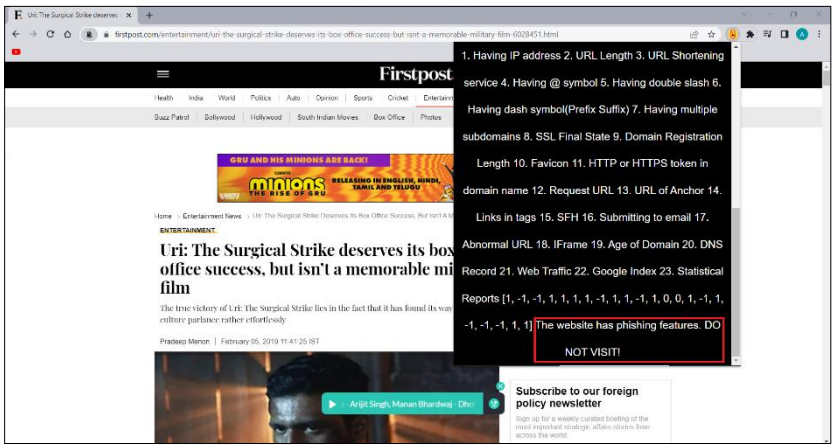

Fig. 4 Phishing Website

VI. Result Analysis

The datasets were split into two parts, Training and Testing, in the ratio 80:20. The training dataset was mainly used to train the different models and then the trained model was applied on the testing dataset to get the results shown in Fig 5.

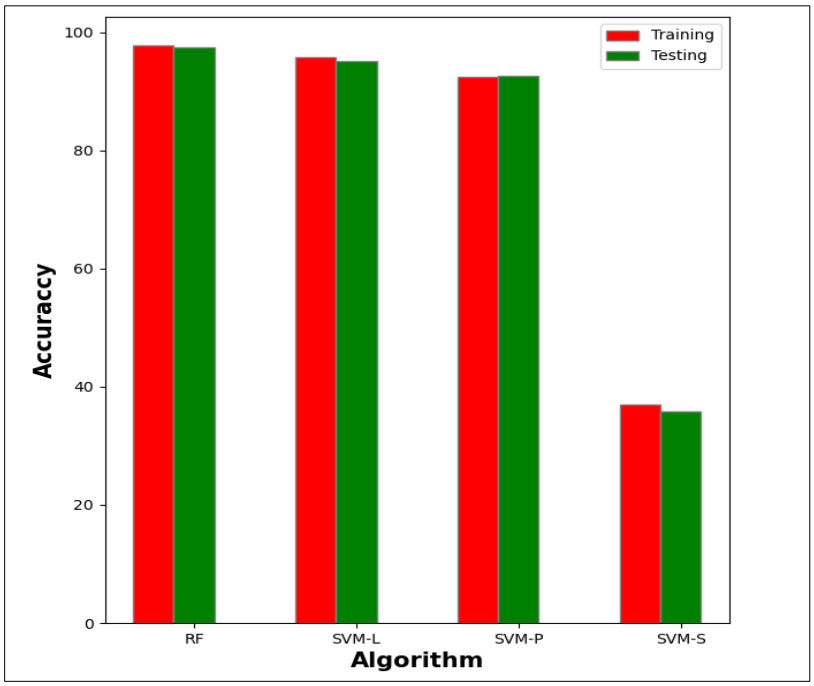

Fig.5 Accuracy obtained from Model

TABLE I Accuracy of Model

Algorithms Used	Accuracy on Training dataset (in %)	Accuracy on Testing dataset (in %)
Random Forest	97.77	97.42
Linear SVM	95.75	95.16
SVM Polynomial Kernel	92.54	92.65
SVM Sigmoid Kernel	37.01	35.83

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified ⁜ Impact Factor 7.39 ⁜ Vol. 11, Issue 6, June 2022
DOI: 10.17148/IJARCCE.2022.11695

VII. CONCLUSION

As a result, we will use Machine Learning to build a prototype model for detecting phishing websites. We're planning to create a system that can quickly identify phishing sites. Python will be utilized as the programming language.

VIII. FUTURE SCOPE

There are billions of people using social media all around the world. However, little is known about how people use social media and the factors that influence how vulnerable they are to phishing attempts on social media platforms. Nonetheless, cyber thieves frequently exploit social networking platforms to defraud their victims. It is a tremendous source of earnings for cybercriminals because there are billions of users signing into their favorite social media accounts. So, after successfully completing this planned job for a bachelor of engineering, we will try to extend our work for social media platforms such as Facebook, Instagram, and others in the future.

REFERENCES

[1]. Atharva Deshpande, Omkar Pedamkar, Nachiket Chaudhary, Dr. Swapna Borde, “Detection of Phishing Websites using Machine Learning”, IJERT 2021.
[2]. Amani Alswailem Bashayr Alabdullah Norah Alrumayh Dr. Aram Alsedrani, “Detecting Phishing Websites Using Machine Learning”, IEEE 2019.
[3]. Alfredo Cuzzocrea, Fabio Martinelli, Francesco Mercaldo, “A Machine-Learning Framework for Supporting Intelligent Web-Phishing Detection and Analysis”, ACM 2019.
[4]. Rishikesh Mahajan “Phishing Website Detection using Machine Learning Algorithms” (2018).
[5]. Purvi Pujara, M. B.Chaudhari “Phishing Website Detection using Machine Learning : A Review” (2018).
[6]. Yasin S¨onmez1 T¨urker Tuncer2 H¨useyin G¨okal 3 Engin Avc4, “Phishing Web Sites Features Classification Based on Extreme Learning Machine.”, IEEE 2018.
[7]. Abdulhamit Subasi, Esraa Molah, Fatin Almkallawi, Touseef J. Chaudhery, “Intelligent Phishing Website Detection using Random Forest Classifier”, International Conference on Electrical and Computing Technologies and Applications (ICECTA), 2017.
[8]. Andrew J. Park Ruhi Naaz Quadari Herbert H. Tsang, “Phishing Website Detection Framework Through Web Scraping and Data Mining”, IEEE 2017.
[9]. M.Phil Scholar, Quaid-E-Millath, “Automated Phishing Website Detection Using URL Features and Machine Learning Technique”, IJET 2016.\
[10]. David G. Dobolyi, Ahmed Abbasi “PhishMonger: A Free and Open-Source Public Archive of Real-World Phishing Websites” (2016)
[11]. Rohan Saraf, Mayur Khatri, Mona Mulchandani “Phish Tank-A Phishing Detection Tool” (2014)
[12]. Satish.S, Suresh Babu.K “Phishing Websites Detection Based on Web Source Code and Url in The Webpage” (2013)
[13]. Matthew Dunlop, Stephen Groat, David Shelly" GoldPhish: Using Images for Content-Based Phishing Analysis” (2010)

