IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Phishing Website Detection Application Using URL Features And Machine Learning

[1]Somalaraju Manoj, [2]Vottikundalu Pavan Kalyan
[3]Dr. A. Gokula Chandar M.Tech., Ph.D.,
SRI VENKATESA PERUMAL COLLEGE OF ENGINEERING AND TECHNOLOGY, PUTTUR

ABSTRACT: Phishing is the simplest method of obtaining sensitive information from unsuspecting consumers. The goal of phishers is to obtain sensitive information such as usernames, passwords, and bank account information. Cyber security professionals are now looking for reliable and consistent detection solutions for phishing websites. This project uses machine learning and databases to detect phishing URLs by extracting and analyzing different aspects of authentic and phishing URLs. The goal of the Paper is to detect phishing URLs and to narrow down the best machine learning method by analyzing each algorithm's accuracy and precision rate. The algorithms are trained using an existing dataset containing of legitimate and phishing URLs and later used to detect phishing websites. This application was developed, keeping in mind the circulation of malicious URLs in social media. Anyone can use this application to detect the validity of the URL before trying to navigate to the website. The use of blacklist approach reduces the response time for phishing detection.

KEYWORDS: *Phishing, Passwords, Cyber security, URLs, Legitimate.*

I. INTRODUCTION

A cyberattack is any offensive tactic that targets computer information systems, computer networks, infrastructures, or personal computer devices. They are unwelcome attempts to steal, expose, alter, disable or destroy information through unauthorized access to computer systems. A cyberattack can be launched from anywhere by any individual or group using one or more various attack strategies. Most common types of cyberattacks are: malware, man-in-themiddle, phishing, distributed denial of service (DDoS), SQL injection, & ran. Web phishing is one of many security threats to web services on the Internet. Web phishing aims to steal private information, such as usernames, passwords, and credit card details, by way of impersonating a legitimate entity. So this project mainly focuses on applying a machine learning framework to detect phishing websites.

II. PROBLEM IDENTIFICATION

With the growth in the field of e-commerce, phishing attack and cybercrimes are rapidly growing. Attackers use websites, emails, and malware to conduct phishing attacks. According to the Anti-Phishing Working Group (APWG) Q4 2020 report, in 2020, there was an average of 225,759 phishing attacks per month, an increase of 220% compared to 2016. The country most affected by phishing sites is China, with 47.9% of machines infected. Phishing has become one of the biggest threats in cyber security. According to the FBI Internet Crime Centre data records, the economic loss due to phishing crimes can reach $3.5 billion in 2019.

Phishing crimes are usually underreported. This technique involves manipulating the user to download malicious files, redirecting to malicious websites, compromising user credentials as well as collecting sensitive information like bank account details etc. Phishing website detection helps the user to determine whether the website is suspicious or not by extracting and analyzing relevant URL features using Blacklist and Machine Learning approach.

III. OBJECTIVES

The main objective of this project is to improve the existing phishing detection techniques. Through this project we aim to:

- Observe the existing and proposed applications for phishing detection.
- Establish the need for incorporating machine learning approach as a solution for the problem statement.
- Collect the datasets of URL-based features for phishing and legitimate websites from reliable open source platforms.
- Identify significant predictors from all available aspects of URL dataset.
- Eliminate duplicate, low-variance data and identify missing values to obtain a valid dataset.
- Choose machine learning algorithms to compare for best classification performance.
- Train classifier using the obtained dataset and consider predictor importance and various evaluation metrics.
- Create a MongoDB blacklist consisting of suspicious URLs.
- Develop feature extraction code to acquire feature values corresponding to the dataset used.
- Develop a UI for user input and display of results.
- Integrate the various components of the project – classifier, blacklist and UI and deploy the final application.
- Help in improvising existing techniques for phishing website detection.

IV. METHODOLOGY

This project proposes a framework that integrates two different approaches for phishing website detection:

1) Blacklist Approach
2) Machine Learning Approach.

The Blacklist method is used for faster detection of illegitimate URLs through quick access into the database. Machine Learning is used for predicting new phishing URLs. The proposed framework comprises of the following steps:

Data Acquisition: This step involves collection of datasets consisting of both phishing and legitimate URLs from open-source platforms like Kaggle. The balanced dataset consists of 11430 URLs with exactly 50% phishing and 50% legitimate URLs.

Data Visualization: Data visualization is the process of analyzing the data and observing how different factors affect the data and how the various columns in the dataset are related with the help of charts, graphs and other visuals. The resulting visual representation of data makes it easier to identify new insights about the information represented in the data. Through this process we are able to observe how each URL feature affects the possibility of the website being phishing.

Database Creation: This step involves creation of blacklist consisting of only phishing URLs. A new csv file is created to store only the illegitimate URLs. This list is then pushed into a mongoDB collection. MongoDB enables the quick access of already stored illegal websites.

Training of Machine Learning Model and Prediction: This process includes selecting the suitable ML algorithm and training it using URL features collected from dataset. 75% of the dataset is used for training. In this project we have used Random Forest Classifier for training the model.

Random Forest Classifier: Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Each tree in a random forest specifies the class prediction, and the result will be the most predicted class among the decision of trees. Random Forests achieve a reduction in overfitting by combining many weak learners that underfit because they only utilize a subset of all training samples. Random Forests can handle

a large number of variables in a data set. Also, during the forest construction process, they make an unbiased estimate of the generalization error. Besides, they can estimate the lost data well. We found that Random Forest is highly accurate, relatively robust against noise and outliers, it is fast, simple to implement and understand, and can do feature selection implicitly. Being unaffected by noise is the main advantage of Random Forest. According to Central Limit Theorem, Random Forest reduces variance by increasing the number of trees

Extraction of URL Features: The user provides URL of the required website as input. Relevant features are extracted from this URL and a dataframe is created. This dataframe is then passed to the trained Random Forest Classifier for prediction. Some of the URL features are:

- **URL Length:** Phishers can use a long URL to hide the doubtful part in the address bar.
- **Having IP Address:** If an IP address is used instead of the domain name in the URL.
- **Shortening Service:** Links to the webpage that has a long URL.
- **Having @ Symbol:** Using the @ symbol in the URL leads the browser to ignore everything preceding the @ symbol and the real address often follows the @ symbol.
- **Prefix Suffix:** Phishers tend to add prefixes or suffixes separated by (-) to the domain name so that users feel that they are dealing with a legitimate webpage.
- **Age of Domain:** If the age of the domain is less than a month.
- **Web Traffic:** This feature measures the popularity of the website by determining the number of visitors.

Automatic Modification of Blacklist: When the URL entered by the user is classified as phishing, it is automatically added to the blacklist database so that in future if the same URL is searched then the system can directly fetch it from database and classify it as phishing.

V. RESULTS

Catch Phish - A Webpage for Phishing Website Detection

We have created a web application called Catch Phish to detect the phishing websites. The user interface is designed using HTML and CSS. Flask framework is used to build the web application.

Homepage: The homepage consists of details about the website. The user can click on „Catch a Phish" button or can click on the „Catch a Phish!" option on navbar to visit the page where he can enter the URL.

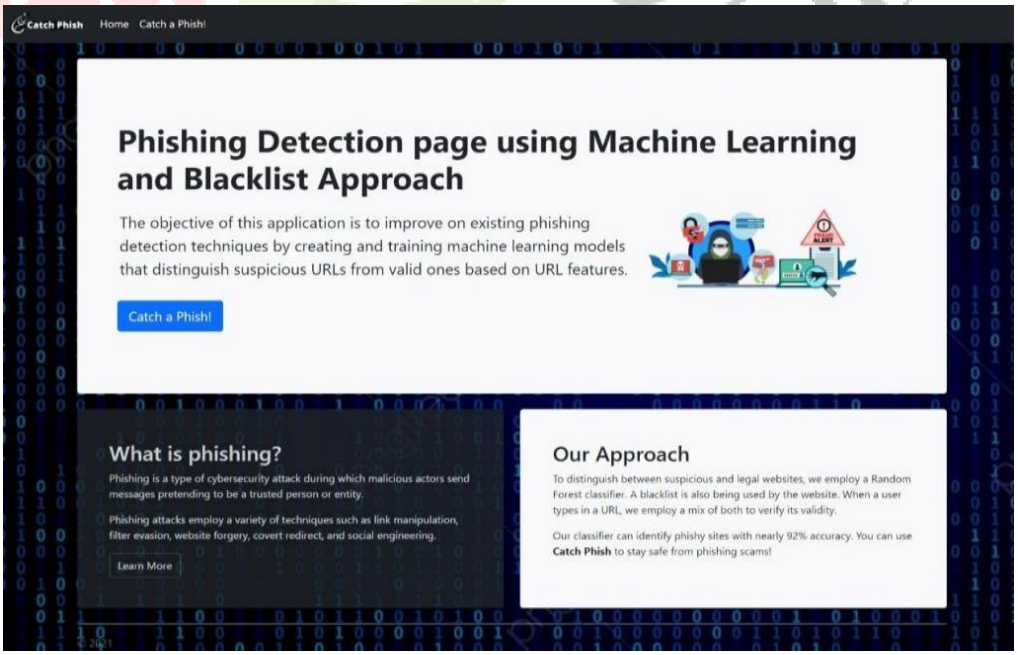

Fig 1: Home Page

Catch a Phish Page: In this page the user can enter the URL of the website he wants to check. After entering the complete URL, the user should click on the „Check for Phish" button. Then the link is sent for verification.

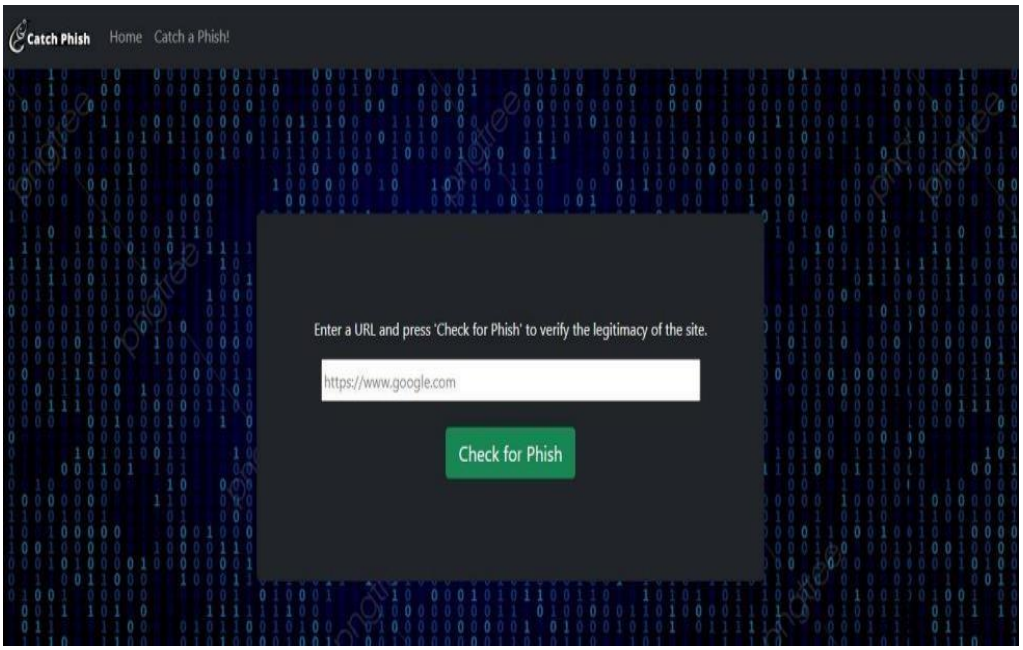

Fig 2: Catch a Phish page

The form will only take complete URL. If the user enters the website name without the https or http token the form will ask the user to enter the complete URL. For example, if the user enters the URL as „www.google.com" the form will ask the user to enter the complete URL including the http or https token.

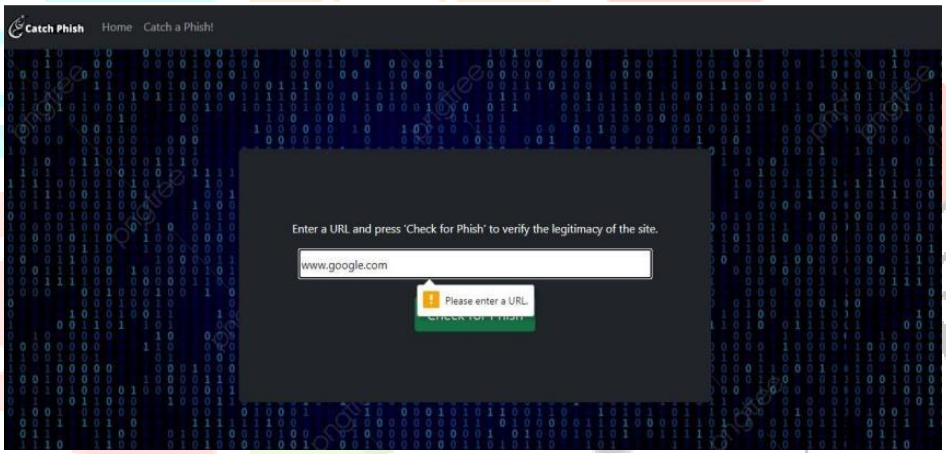

Fig 3: Invalid URL Entry

Result Page: If the user enters the valid URL, it will be verified and the user will be sent to the result page where the application will display whether the URL is fake or legitimate. If the entered URL is classified as legitimate then the webpage will notify the user about its legitimacy.

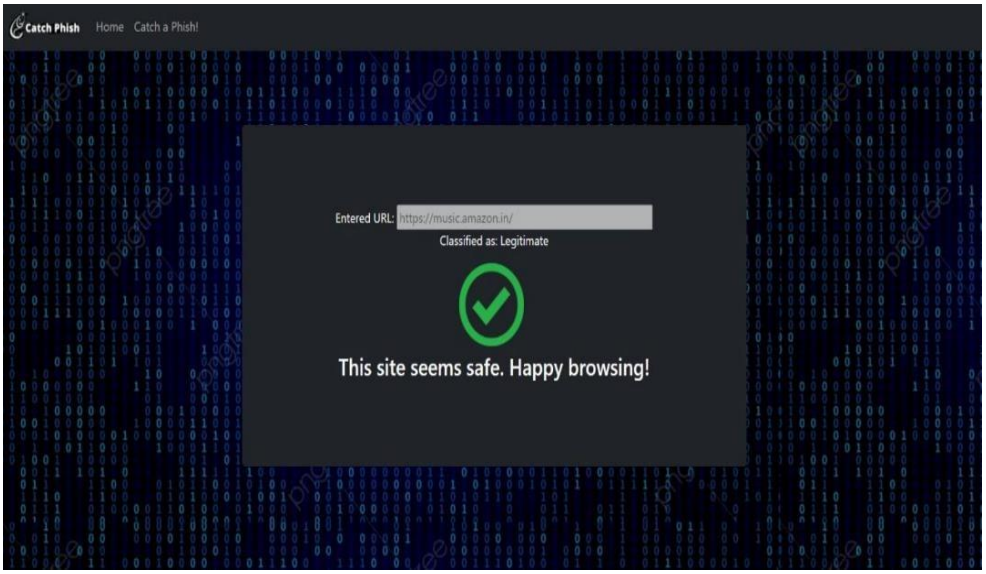

Fig 4: Results for Legitimate URL

If the URL is classified as phishing then the system will warn the user to not to visit the fake website.

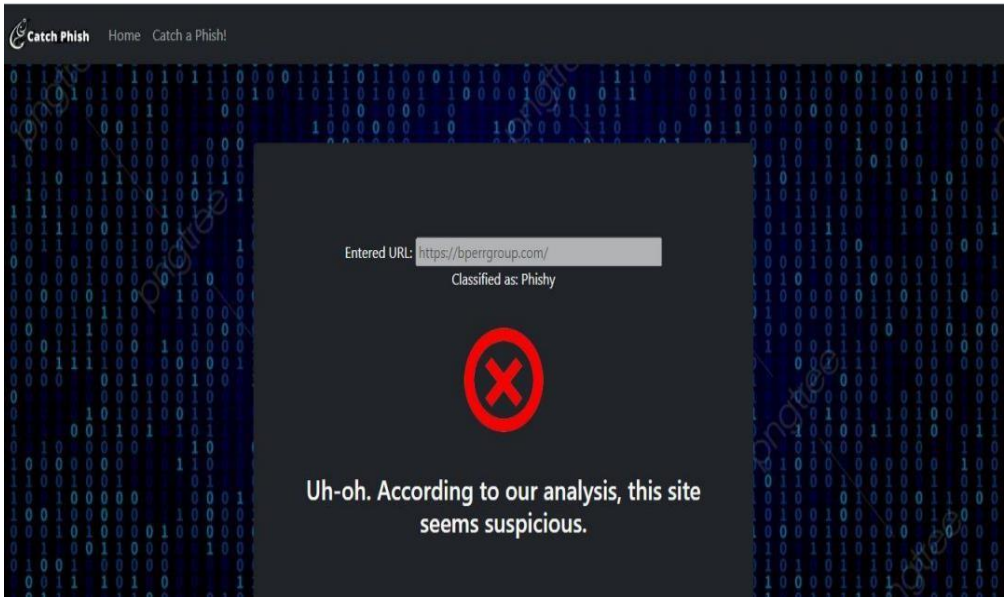

Fig 5: Result for phishing URL

VI. CONCLUSION

Through this project, we have improvised the current phishing detection techniques by combining both blacklist and machine learning approaches. The use of blacklist approach reduces the response time since the URL will be searched in the database before going through feature extraction and being passed to the classifier for prediction. We have considered 22 URL features for which we obtained the accuracy up to 92%. This application can be used by everyone to detect the phishing websites which are circulated nowadays through social medias. The simple design helps the user to understand the usage and easily access the application. This can be helpful in reducing the phishing attacks which are being carried out by sharing the malicious URLs.

VII. REFERENCES

[1]. Gokula Chandar A, Vijayabhasker R., and Palaniswami S, "MAMRN – MIMO antenna magnetic field", Journal of Electrical Engineering, vol.19, 2019.

[2]. Rukkumani V , Moorthy V, Karthik M , Gokulachandar A, Saravanakumar M, Ananthi P, "Depiction of Structural Properties of Chromium Doped SnO2 Nano Particles for sram Cell Applications", Journal ofMaterials Today:
Proceedings, vol.45, pp.3483-3487, 2021. https://doi.org/10.1016/j.matpr.2020.12.944

[3]. Chandar AG, Vijayabhasker R., and Palaniswami S, "ILAPARC-Isolation mimo LTE Antenna Placement in Wireless Devices with Adjustable Radiation Control" ,Journal of tierarztilich praxis, vol.39, no.11, 2019.

[4]. Gokula Chandar ,Leeban Moses M; T. Perarasi M; Rajkumar; "Joint Energy and QoS-Aware Cross-layer Uplink resource allocation for M2M data aggregation over LTE-A Networks", IEEE explore,doi:10.1109/ICAIS53314.2022.9742763.

[5]. Dhuddu Haripriya, Venkatakiran S, Gokulachandar A, "UWB-Mimo antenna of high isolation two elements with wlan single band-notched behavior using roger material",Vol 62, Part 4, 2022, Pg 1717-1721, https://doi.org/10.1016/j.matpr.2021.12.203.

[6] Jain, A.K.; Gupta, B. Comparative analysis of features based machine learning approaches for phishing detection;
pp. 2125–2130

[7] Lee, L.H.; Lee, K.C.; Chen, H.H.; Tseng, Y.H. Poster: Proactive blacklist update for antiphishing; pp. 1448–1450.

[8] Rao, R.S.; Ali, S.T. Phishshield: A desktop application to detect phishing webpages through heuristic approach. Procedia Comput. Sci. 2015, 54, 147–156.

[9] Zhang, Y.; Hong, J.I.; Cranor, L.F. Cantina: A content-based approach to detecting phishing web sites;

pp. 639– 648.
[10] Mohammad, R.M.; Thabtah, F.; McCluskey, L. Predicting phishing websites based on selfstructuring neural network. Neural Comput. Appl. 2014, 25, 443–458.
[11] Lin, Y.; Liu, R.; Divakaran, D.M.; Ng, J.Y.; Chan, Q.Z.; Lu, Y.; Si, Y.; Zhang, F.; Dong, J.S. Phishpedia: A Hybrid Deep Learning Based Approach to Visually Identify Phishing Webpages.
[12] www.kaggle.com
[13] https://archive.ics.uci.edu/ml/datasets/phishing+website
[14]https://www.researchgate.net/publication/328541785_Phishing_Website_Detection _using _Machine_Learning_Algorithms
[15] Varshney G, Misra M, Atrey PK. A phish detector using lightweight search features. Computers & Security 2016; 62:213-228
[16] Xiang, G.; Hong, J.; Rose, C.P.; Cranor, L. Cantina+ a feature-rich machine learning framework for detecting phishing web sites.ACM Trans. Inf. Syst. Secur. 2011, 14, 1–28.
[17] Lee, L.H.; Lee, K.C.; Chen, H.H.; Tseng, Y.H. Poster: Proactive blacklist update for antiphishing. In Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, 3–7 November 2014; pp. 1448–1450.
[18] Chiew KL, Chang HH, Sze SN, Tiong WK. Utilisation of website logo for phishing detection. Computers & Security 2015 ; 54:16-26.
[19] Hadi, W.; Aburub, F.; Alhawari, S. A new fast associative classification algorithm for detecting phishing websites. Appl. Soft Comput. 2016, 48, 729–734

BIBLIOGRAPHY

I am **SOMALARAJU MANOJ**, Growing up I was always interested in the Computers and technology. My passion for ECE continued to evolve during my undergraduate studies. I later discovered the transformative power of data in technology through machine learning and deep learning. This fusion of ECE and data science solidified my interest.

I am **VOTTIKUNDALU PAVAN KALYAN**, When I was younger I was passionate about Engineering, which prompted me to pursue a bachelor‟s degree in Sri Venkatesa Perumal College of Engineering & Technology, Puttur. During college, I am a person who has always had a profound passion and fascination for areas requiring an analytical approach and quantitative thinking.