iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 8 **Issue:** VII **Month of publication:** July 2020

**DOI:** https://doi.org/10.22214/ijraset.2020.30421

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429*
*Volume 8 Issue VII July 2020- Available at www.ijraset.com*

# Phish-Hook: Phishing Site Detection using URL Features

LaxmiPriya Nadar[1], Hema Thakur[2], Poonam Yadav[3], Dipali Junankar[4]
[1, 2, 3]*Student, Computer Engineering, New Horizon Institute of technology & Management (Thane), Maharashtra, India*
[4]*Professor, Computer Engineering, New Horizon Institute of technology & Management (Thane), Maharashtra, India*

***Abstract:*** ***This project aims at detecting phishing sites made by the phishers who steal personal user data to conduct illicit activities. We will extract features from the URL which are submitted by the user. These features will then be given to decision tree algorithm to classify the site as phishing or legitimate. Also, ranking of the sites will be considered while classifying the site as phishing or legitimate. In this project URL- based heuristic approach will be used along with the ranking of sites to extract the features from the URL. All the extracted features along with the phishing and legitimate sites URLs will be stored in database. Next, a classifier will be generated using decision tree algorithm which will classify the URLs as phishing and legitimate. When new URL is received it will extract the features and will compare them with the features stored in database, thus classifying the incoming site as phishing or legitimate.***
***Keywords: GUI, Decision Tree Algorithm, Random Forest Algorithm, Page Rank, Alexa Rank, Dmoz, Phishtank.***

## I. INTRODUCTION

Phishing is a malicious use of Internet resources carried out to trick Internet users to reveal personal information, such as usernames, credit card information, and Social Security numbers to the attacker and for fraud. Phishing can appear through a variety of communication forms such as instant messaging, SMS, VOIP, online messenger, and above all the most common form of phishing attack leverages email. Fraudsters send an email to an unsuspecting user that contains a link to a domain that is seemingly legitimate in the hopes that the users will input their private information for the attacker to steal. There is no doubt that phishing can be extremely damaging all organizations over internet since tricking a user within a business network through a phishing scam is very easy way to obtain the user's information in order to gain access to that business network and many more..

Phishing can have a large impact on individual Internet users. According to the APWG Report, among the top-level domains the .COM namespace contained the most unique domain names used for phishing as well as having the highest number of phishing attacks within the namespace in the quarter of year 2013 . This would suggest that a large number of phishing attacks targeted typical Internet users and not corporations.

## II. RELATED WORK

The phishing technique aims to fool the online users by making a fake URL which is similar to the URL of the original website, the domain-related features of the URL can be used to detect the phished websites. Specifically, Primary Domain, Sub Domain and Path Domain of the URL are investigated to conclude the websites. Also, the ranking of site such as Page Rank, AlexaRank can also help to detect phishing sites. It is included in the heuristic set to achieve higher accuracy detection level.[1] [2]

Machine learning algorithms are used to build an efficient classifier which would decide whether a given URL is phishing or not. . It creates a tree form for classifying samples. Each internal node of the tree corresponds to a feature, and the edges from the node separate the data based on the value. Decision tree includes a decision area and leaf node. The decision area checks the condition of the samples And separates them into each leaf node or the next decision area.The decision tree is very fast and easy to implement.[3][4]

Zhang et al presented a tool called Cantina in 2007. CANTINA examines the content of a web page to determine whether it is legitimate or not, in contrast to other approaches that look at surface characteristics of a webpage, for example the URL and its domain name. CANTINA makes use of the well-known TF-IDF (term frequency/inverse document frequency) algorithm used in information retrieval. CANTINA combined with heuristics is effective at detecting phishing URLs in user's actual email, and that its most frequent mistake is labeling spam-related URLs as phishing.[4][5] Random forest is a classification method that combines many tree predictors; each tree depends on the values of a random vector that is independently sampled. All trees in the forest have the same distribution. This algorithm can handle a large number of variables in the dataset, however, it lacks reproducibility because the process of forest building is random ,hence gives low accuracy.

## III. SYSTEM DESIGN

URL: A URL (uniform resource locator) is used to locate the resources. An example of a typical URL would be ”http : //en.example.org/wiki/MainPage”. However, in many of the cases, URL is often used as a synonym for URI .
The structure of URL is as follows: < protocol >: // < subdomain > . < primarydomain > . < TLD > / < pathdomain >
For example, the URL: http://www.paypal.attack.com/login/index.php, there are six components as follows: Protocol is http, Subdomain is paypal, Primarydomain is attack, TLD is com, Domain is attack.com and Pathdomain is login/index.php.

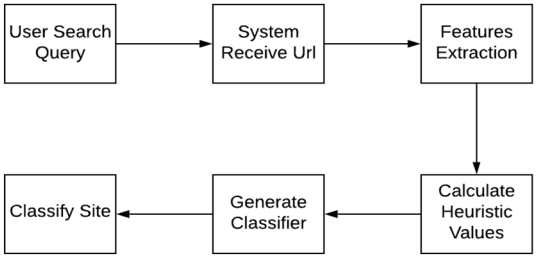



Fig. Block Diagram

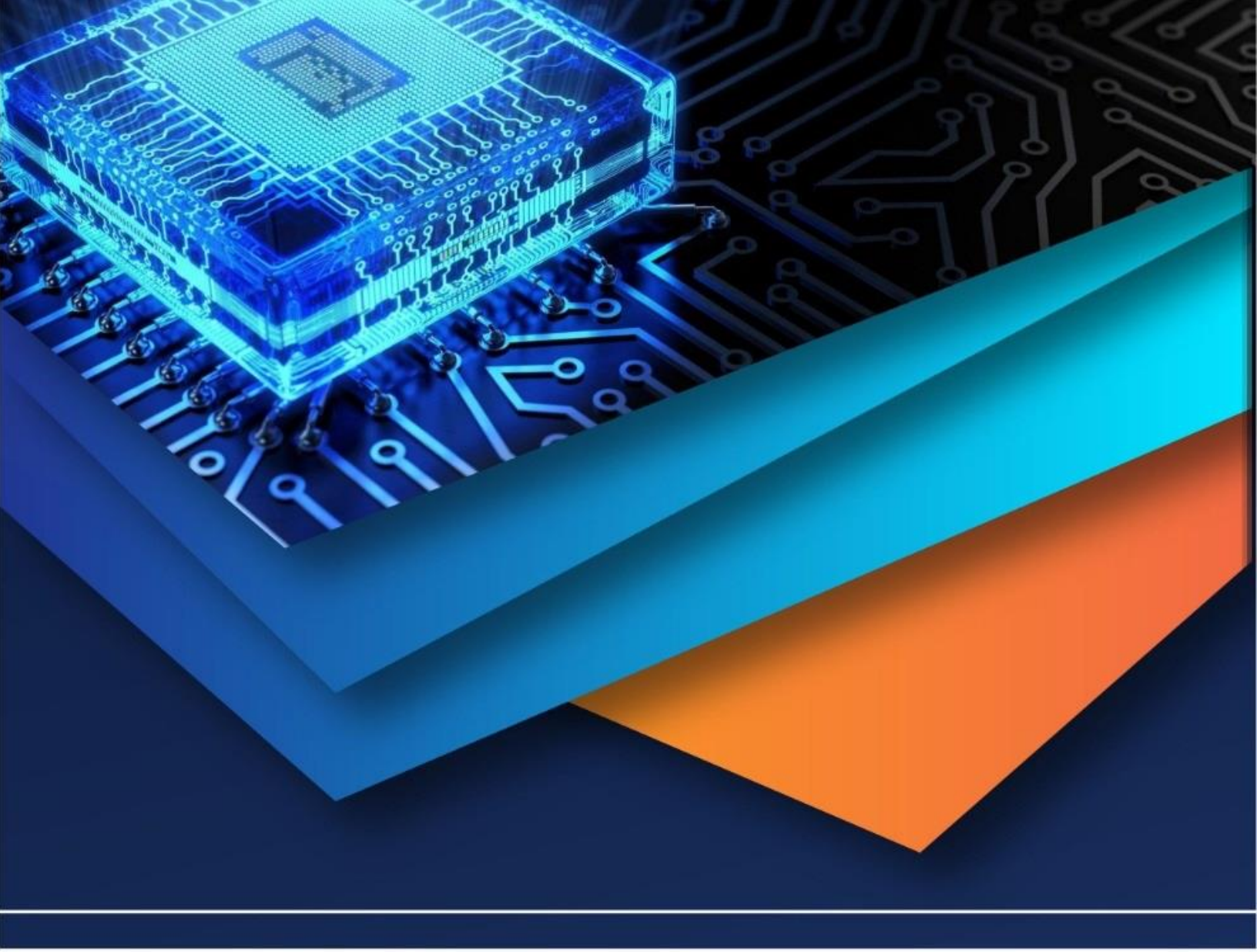
The training set works as follows:
Initially, the user will enter his/her URL. This will be in the complete format
E.g.: https://www.google.com. Then after clicking enter option the system will receive the URL. Here the system will search for the entered URL in the database on the list consisting of phishing and legitimate sites. If the URL is present in the list then the system will proceed for providing the final output or if the provided URL is not present in the list the process of feature extraction will take place. Here the URL will be divided into domains and sub domains. Basically, The URL is composed of the protocol, sub domain, primary domain, top-level domain (TLD) and path domain. Protocols are of various types and are used in accordance with the desired communication method. The sub domain is an ancillary domain given to the domain and has various types depending on the services provided by the domain page. The domain is the name given to the real Internet Protocol (IP) address through the Domain Name System (DNS). The primary domain is the most important part of a domain. The TLD is the domain in the highest position in the domain name hierarchy architecture.e.g. .com, .net, .kr, .jp, etc. We define features of each component of the URL these features are used for phishing site detection.
We use the formula for calculating the heuristic value. If the heuristic value results as 1 then it are a phishing site else the site is legitimate. The final classifier is generated then stored in database and the resultant output will be given in the form of length of http, http present or not suspicious character found or not, number of dots, length of subdomain, number of slash present in the URL.

## IV. CONCLUSION

We generated classifiers through several machine learning algorithms and determined that the best classifier was ID3. It showed a high accuracy of 73% and a low false-positive rate. The proposed technique can provide security for personal information and reduce damage caused by phishing attacks because it can detect new and temporary phishing sites that evade existing phishing detection technique.

## REFERENCES

[1] Xiaoxin Yin, Jiawei Han, Senior Member, IEEE, and Philip S. Yu, Fellow, IEEE, “Truth Discovery with The Multiple Conflicting Information Providers on the Web”, Los Angeles, CA, USA, VOL. 20, NO. 6, JUNE 2008, pp. 796-808
[2] Xin Luna Dong, Laure BertiEquille, Divesh Srivastava, “Truth Discovery and Copying Detection in a Dynamic World”, VLDB „09, August 2428, 2009.
[3] Gil, Y. and Artz, D.,”Towards content trust of the web resources”, Edinburgh, Scotland, May 23 - 26, 2006, DOI=http://doi.acm.org/10.1145/1135777.1135861 ,NY, pp565-574.
[4] Soo Young Rieh,”Judgment of Information Quality and Cognitive Authority in the Web”, citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.107.8991
[5] Nguyen, Luong Anh Tuan, et al.A novel approach for phishing detection using URL-based heuristic Computing, Management and Telecommunications (ComManTel), 2014 International Conference on. IEEE, 2014.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY