International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 11 | Nov 2022 www.irjet.net p-ISSN: 2395-0072

# PHISHING URL DETECTION USING MACHINE LEARNING

**[1]Prof. Pooja Parikh,[2]Ketan Kokane, [3]Shrikant Rathod, [4]Mayur Mali,[5]Harshal Pagare**

*ALARD COLLEGE OF ENGINEERING & MANAGEMENT*
*(Alard Knowledge Park, Survey No. 50, Marunji, Near Rajiv Gandhi IT Park, Hinjewadi, Pune-411057) Approved by AICTE. Recognized by DTE. NAAC Accredited. Affiliated to SPPU (Pune University)*

---

## I. ABSTRACT:

Phishing is an illegal activity that uses a variety of deceptive methods to direct people to the wrong website. The purpose of these phishing websites is to confiscate personal information and other financial details for personal gain or abuse. As technology advances, the phishing approaches in use must evolve, and there is an urgent need for increased security and improved mechanisms to prevent and detect these phishing approaches. The main focus of this paper is to introduce his model as a solution for detecting phishing websites using the URL detection method with a random forest algorithm. The model has three main phases, such as parsing, heuristic classification of data, and performance analysis, where each phase uses a different technique or algorithm to process the data for better results.

## II. INTRODUCTION

Online procedures, online business or trading, or exposure, so the online systems already in place at that time faced little threat. However, in the past five years, the world has experienced a big boom in the IT sector, resulting in most of the daily operations going online. From shopping to banking. The term "phishing" was coined in 1996 by his hacker, who stole the America On-Line account by stealing passwords from unsuspecting AOL users. The word phishing comes from the phrase "website phishing" and is his variation of the word "phishing". The idea is that, like a fish, it casts the bait in hopes that the user will grab it and bite. In most cases, bait is either an e-mail or an instant messaging site, which will take the user to hostile phishing websites. Over the years, phishing attacks grew in number and intensity too.

Phishing attacks now target users of online banking, payment services such as PayPal, and online e-commerce sites. There are different modes through which phishing can be carried out and hence there are various types of phishing like vishing (voice over phishing), smishing (Phishing via SMS), whaling, Mishing (mobile phishing), social engineering, spear phishing, etc. Usually, there are four phases in a typical phishing attack like preparation, mass broadcast, mature and account hijack. For most of the phishing attacks, whether carried out by emails or any other medium, the objective is to get the victim to follow a link that appears to go to a legitimate web resource but actually redirects the victim to a malicious web page. The easiest way to link the operation is to construct a malicious URL and direct the user to the malicious page desired by the attacker. This document focuses on detecting phishing websites with URL detection. Previously, methods such as k-nearest neighbors, list-based approaches, fuzzy logic, mining, and classification approaches such as Phishzoo were used for detection, but over time as the strength of attacks increased , a more sophisticated algorithm was used. techniques were introduced to detect and prevent attacks.

## III. PREVIOUS WORK

Currently, various peer-reviewed journals and conferences have published various studies and studies on phishing website detection, and one proposed approach was multi-level classification of phishing URL filtering. In it, author presents an innovative method for extracting phishing URL features based on message content weighting [3]. His multiple classification algorithms including SVM, AdaBoost and Naive Bayes are used. These algorithms are divided into three layers using 21 fixed individual functions 4446 [3]. A two-step process is then done using a different classification algorithm, but the problem here is the time and complexity involved, the overhead 4486 involved, and the performance issue, so this is It wasn't the best 4484 method.

Another method adopted by one author of the IEEE 2017 paper was pattern recognition, d. H. Various features are extracted from emails to obtain a model that helps distinguish between phishing and non-phishing messages [4]. One of the main methods used in this context is detecting attacks and using feature extraction and classification. The main limitation of this proposal is that it evaluates too many characteristics without considering whether they are really important for identifying phishing. Therefore, this can lead to unnecessary computational costs. According to the Institute of Research Engineers and Doctors, USA,

phishing detection techniques can be divided into blacklist-based and heuristic-based approaches [5]. A blacklist-based approach maintains a database listing the addresses (URLs) of sites classified as malicious. If a user requests a site with in this list, the connection will be blocked. The blacklist-based approach has the advantages of simple implementation and low false positive rate [5]. However, has a bug that prevents it from detecting phishing sites not listed in the database, including temporary sites.

According to the International Journal of Advanced Research and Innovative Ideas in Education (IJARIIE) Journal, a data mining approach based on the Multi-Label Classifier Association (MCAC) is also one of the methods used to detect phishing websites . The associative classification algorithm detects phishing websites with moderate accuracy. [6] MCAC consists of three main steps: rule recognition, classifier formation, and class assignment. In the first step of this algorithm, rules are detected and extracted by iterating over a training dataset (historical website features or data collected from various sources). that have the same precondition (left side) and are associated with different classes to create a multi-label rule. Along with, redundant rules are eliminated. The result of the second step is a classifier containing single and multiple labeling rules. The final step is to test the classifier against the test dataset to measure its performance. In the prediction process, rules in the classifier set that match test data features are often triggered to infer their type (class). The MCAC algorithm also generates rules that are sorted using the sorting algorithm. [6]. The main problem faced by MCAC was the difficulty of determining minimum confidence and support in the presence of large amounts of data. Later, these more accurate and less time-complex algorithms were replaced by more sophisticated algorithms.

## PROPOSED SYSTEM

Our proposal for the above topic is to improve the detection efficiency of phishing websites. A number of surveys and polls were conducted to compare different classification algorithms to best fit our model [7]. In addition to using WEKA to determine the accuracy and performance of each algorithm, many journals and articles were read and researched to determine classification algorithms. The idea put forward here is to improve efficiency by using random forests as classification algorithms with the help of Rstudio tools that help in better analysis. Below is a flowchart of our proposal.

| | |
|---|---|
| 1 | IP Address |
| 2 | Redirection of page using "//" |
| 3 | Adding Prefix or Suffix Separated by (-) to the Domain |
| 4 | Sub domain and multi-sub domain |
| 5 | URLs having @ symbol |
| 6 | Using different functions in the URL to submit information |
| 7 | Page Rank |
| 8 | Google Index |

The datasets required for the entire procedure were obtained from phishing tanks [8], and the analysis of was mainly performed due to the large amount of data that had to be processed. analyzes are performed to analyze the feature set. We restrict feature set to 8 of the 31 features considered by parsing and rigorous analysis of, which are shown in Table 1. where parsing is done using attribute subset selectors. It consists of two parts: 1) Attribute Subset Scoring Algorithm using Information Gain 2) Search Method Algorithm using Ranker Method. Parsing is implemented in Java code that imports the IG and the WEKA tools library for attribute selectors. Attributes that provide more information have higher information gain values and can be selected, while attributes that do not add much information will score lower and can be removed. Table 1 shows the features we included in our model to help with classification.

Since the entropy of the training set S is the impurity criterion, we can define a measure in weka that reflects the additional information about Y provided by X representing the amount by which the entropy of Y is reduced [9]. IG is given by

formula: IG = H(Y) − H (Y \X) = H (X) − H (X \Y) (1)

IG is a symmetrical scale. measure. The information about Y after observing X is the same information about X after observing Y. The parsed records undergo a heuristic classification splitting the records into 70% and 30%. 70% of the data is considered for training and 30% for testing. Using the random forest library and built-in R functions, a classification model 4484 is built and tested using test data. This model is used to predict other URLs for various websites entered by the user. The final phase of the model run is a performance analysis performed using the ROC curve. In addition to the ROC curve, other factors such as sensitivity, confusion matrix and fp rate are also included.

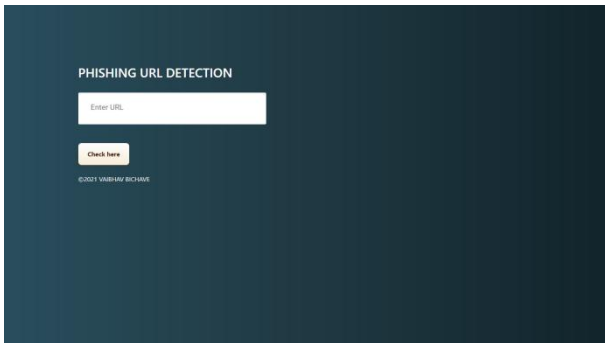


Fig 2(a)

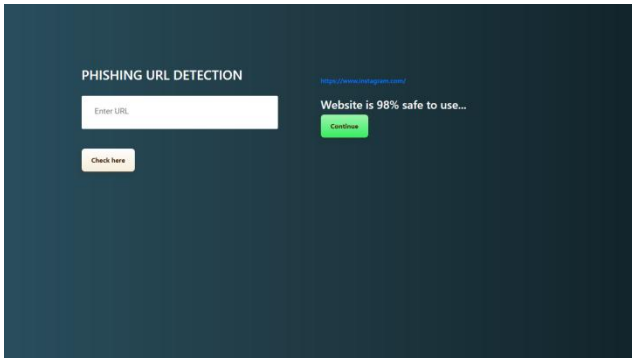


Fig 2(b)

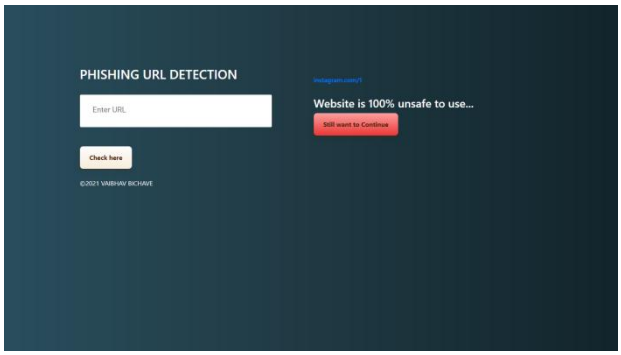


Fig 2(c)

## CONCLUSION

In this document, with the help of Rstudio, I proposed another method to detect phishing websites using Random Forest as a classification algorithm. Here, we have empirically shown that 31 of the proposed features are the best for detecting phishing websites. Random forests were chosen for classification because the performance metrics and our literature review also proved that random forests had the highest level of accuracy, around 95% [10]. The model uses a wide range of metrics such as true positives, true negatives, false negatives, F value, ROC, accuracy and sensitivity for analytical purposes, clearly demonstrating the performance and accuracy of each detection. There is currently no one-size-fits-all solution to phishing, and future technologies are expected to increase the types and numbers of phishing attacks. To do this, the browser must be able to configure how it detects and warns of potential phishing attacks. Future work aims to add improved features to the detection process to develop a system that can learn by itself about new types of phishing attacks.

## ACKNOWLEDGEMENT:

We put a lot of time and effort into this task. However, many individuals have helped and assisted us in completing this task. I would like to express my sincere gratitude to everyone at

We thank Prof Pooja Parikh for his role as our guide and for always being there by keeping an eye on us and providing his necessary input and information on the project. Thanks to Prof Priyardarshini Doke.

We thank the parents and friends of Gargotta's University for their encouragement and cooperation. Industry employees thank you very much for your attention and time.

## REFERENCES

[1] APWG-Unifying Global response to cybercrime http://docs.apwg.org/word_phish.html

[2] International Journal of Advanced Research in Computer(IJARCET) – "An Efficient Approach To Detecting Phishing A Web Using K-Means And Naïve-Bayes Algorithms"

[3] International Journal of Advanced Computer Technology (IJACT), "A Review of Various Techniques for Detection and Prevention for Phishing Attack".

[4] IEEE 2017 - Feature selection for machine learning based detection of phsihng websites "http://ieeexplore.ieee.org/abstract/document/8090317/?reload=true"

[5] Heuristic-based Approach for Phishing Site Detection Using URL Features- Third Intl. Conf. on Advances in Computing, Electronics and Electrical Technology - CEET 2015 Copyright © "Institute of Research Engineers and Doctors, USA. All rights reserved."

[6] Prof.T.BhaskarAher Sonali, Bawake Nikita , Gosavi Akshada ,Gunjal Swati, ' Detection of Website Phishing Using MCAC Technique Implementation', "http://ijariie.com/AdminUploadPdf/Detection_of_Website_Phishing

_Using_MCAC_Technique_Implementation_ijariie1807.pdf"

[7]2015 9th International Conference on Software, Knowledge, Information Management and Applications (SKIMA) , "Detection of Phishing Emails using Data Mining Algorithms"

[8] Phishtank- "https://www.phishtank.com"

[9] "Performance Comparison of Feature Selection Methods" -Thu Zar Phyu , Nyein Nyein Oo, MATEC Web of Conferences.

[10]A Novel Multi-Layer Heuristic Model for Anti-Phishing- "https://dl.acm.org/citation.cfm?id=3078580", ACM", paper 2017.