# Analysis on Suitable Machine Learning Model for Fraudulent Website Categorization

[1] Sk .Jakeer Shareef, [2] G .Anil Kumar, [3] T .Praveena, [4]Pinnelli. Srija
[1,2,3]Assistant Professor, [4]Student, Dept. of Computer Science Engineering, Newton's Institute of Engineering, Macherla, Andhra Pradesh, India.

**Abstract**
Phishing attacks are a type of cybercrime that are made successful by malware and social engineering. It is a serious danger that must be accepted by everyone and everything. Uniform Resource Locators (URLs), sometimes referred to as web links, are the mechanism by which people search the internet for particular pages. The analysis educates readers about phishing, shows them how to spot phishing attempts, and inspires them to take preventative measures. Phishers trick their victims by sending emails or messages that contain links to dangerous websites (known as "phishing" or "spear phishing"). Businesses and individuals are unable to identify all of the phishing emails and messages that are sent to them because of the sheer number that they receive daily. We explore numerous machine learning techniques for phishing attack detection in this section. It is used here to determine whether a particular group of links is a phishing effort or not. Phishing is a common strategy used by cybercriminals to deceive victims into providing personal information. Phishing web addresses are intended to steal sensitive information such as login passwords and financial data from unwary users. Phishing websites can sometimes closely mimic authentic ones, both visually and philosophically. Anti-phishing technologies are required because phishing attempts are becoming more sophisticated as technology advances. Machine learning has emerged as a powerful tool in the fight against phishing. This study gives an in-depth look at machine learning detection techniques and the features they employ.
**Keywords:** URL, PHSHING SITES, Random Forest, XGBOOST

## Introduction

One of the most deadly types of cybercrime is phishing. Due to the large number of people who use the internet to access official government and financial services during the past several years, phishing attacks have increased dramatically. There are now profitable phishing operations on the rise. To deceive unwary users into disclosing their personal information, phishers employ a broad range of strategies, including email, instant messaging, and voice over IP, forged links, and bogus websites. It's easy to create fake websites that replicate the look and feel of authentic ones. Furthermore, the information on these websites would be identical to that on their legitimate equivalents. These websites exist solely to steal users' private data, such as passwords, login IDs, and debit and credit card details. Attackers also pose as high-level security personnel and ask users to answer security questions. Users are readily duped into falling for phishing schemes when they respond to those inquiries. There have been many efforts from many communities all around the world to find ways to stop phishing assaults. The best way to stop phishing assaults is to be informed of how to spot fake sites and spread that knowledge to the general public. One of the effective methods for identifying phishing websites is the use of machine learning algorithms. In this research, we examined many strategies for identifying potential phishing sites. Social engineering remains one of the simplest and most effective techniques to obtain access to private information, despite the increasing volume and sophistication of cybersecurity threats. According to the United States Computer Emergency Readiness Team (US-CERT), "phishing" is "a form of social engineering in which an individual or organisation fraudulently presents themselves as a legitimate one over an electronic communication channel in order to obtain sensitive information from the recipient" [1]. While it is important for businesses to teach their staff how to spot phishing emails and URLs in order to protect themselves from the aforementioned threats, consumers can easily make copies of entire websites with tools like HTTrack. Therefore, even

knowledgeable users might be duped into providing critical information to a malicious website that appears to be legitimate. In light of the aforementioned issue, it is clear that both user education and technological safeguards against phishing attempts are required. The capacity of a computer to recognise potentially harmful websites would be crucial in preventing its users from accessing such sites. Uniform Resource Identifiers (URLs) are one method of identifying phishing websites that are not real (URLs). To find a certain web page or resource on the Internet, you need to know its Uniform Resource Locator (URL). However, the URLs can be utilised to tell the difference between the original and fake sites even if the content is identical. Anti-virus organisations have created a blacklist of harmful URLs as one possible solution. The issue with this strategy is that new malicious URLs appear constantly, making itimpossible to create an extensive blacklist. Therefore, methods are required that can immediately identify if a novel, unseen URL is a phishing site or not. These methods often rely on machine learning to classify new phishing sites based on a model created from existing attack data.

**Literature survey**

The majority of current machine learning approaches for detecting phishing URLs use some extracted feature or group of features to evaluate a URL. Extraction of features from URLs can be broken down into two broad categories: host-based features and lexical features. Information about the website's host, such as its location, administrator, and installation date, is known as "host-based features." Lexical attributes, on the other hand, characterise the URL's textual characteristics. Since Uniform Resource Locators (URLs) are just strings oftext that can be broken down into their component parts—the protocol, hostname, and path—a system can evaluate a website's reliability using any of those factors.

The detection of malicious URLs has been the subject of numerous machine learning approaches. To organise potential phishing websites, Sadeh et al. [2] suggested the PILFER system. They culled eight characteristics that were developed to expose tricksters' dishonest strategies. Roughly 860 legitimate emails and 6950 phishing emails make up the dataset. A Support Vector Machine (SVM) was employed to make classifications in the final product. Using 10-fold cross validation for training and testing, they achieved 92% accuracy. To solvethis issue, Ma et al. [3] modelled the URL classification problem as a binary classification problem and developed a URL classification system that takes in a stream of labelled URLsin real time. It also gathers real-time URL characteristics from a major web-based email service. They made advantage of host-based properties in addition to lexical ones. In order to train an online classifier using the collected data and labels, they used a Confidence Weighted(CW) approach. After reviewing 358 studies, Parkait et al. [4] give a thorough literature analysis on the topic of phishing counter measures and their efficacy. They categorised phishing-prevention techniques into eight distinct categories and emphasised cutting-edge phishing-prevention strategies. Multi-label Classifier based on Associative Classification is a technique developed by Abdelhamid et al. [5] with the purpose of identifying phishing URLs (MCAC). Using these sixteen features, they were able to place URLs into one of three categories: phishing, legitimate, and suspect. As a rule-based algorithm, the MCAC uses the phishing dataset to derive numerous label rules. Within their assessment of methods for detecting malicious webpages, Patil and Patil [6] provide a high-level summary of several types of web-page attacks. Fast-Associative Classification Algorithm (FACA) was used by Hadi et al. [7] to categorise potential phishing URLs. FACA is effective because it constructs a model for categorization by uncovering all sets of frequently occurring rules. They looked into a dataset that contained 11,055 sites classified as either real or phishing. Thirty characteristics were included in the dataset. At least a 50% level of confidence was required, and 2% of support was the bare minimum employed. To identify harmful URLs, Nepali and Wang [8] introduced a new method that relies solely on publicly available information from social media. With the help of supervised learning and features points gleaned from WHOIS and DNS data, Kuyama et al. [9] suggested a method for determining which server acts as the Command and Control (C&C) node. They used the WHOIS information for evaluatingdomain names and emails. Several scholars have also conducted surveys of the topic of malicious URL detection, adding to the aforementioned options. When it comes to machine learning for detecting malicious URLs, Sahoo et al. [10] present a thorough overview and a structural understanding.

**Proposed system**

The experimental framework, ML algorithms, data collection, and performance metrics areall discussed in this section. The conceptual diagram of the suggested experiment is depicted in Figure 1. First, a database of phishing websites is chosen. After that, features are retrieved and the data is cleaned up to make it more manageable. Then, ML classifiers are fed the features derived by PCA (PCA). The most efficient algorithms are stacked together to achieve the desired result. Ten-fold cross-validation is used to train the features. Classifiers' efficacy in spotting phishing websites can be measured using these evaluation metrics. Fig. 1 depicts the overall structure that will be used.

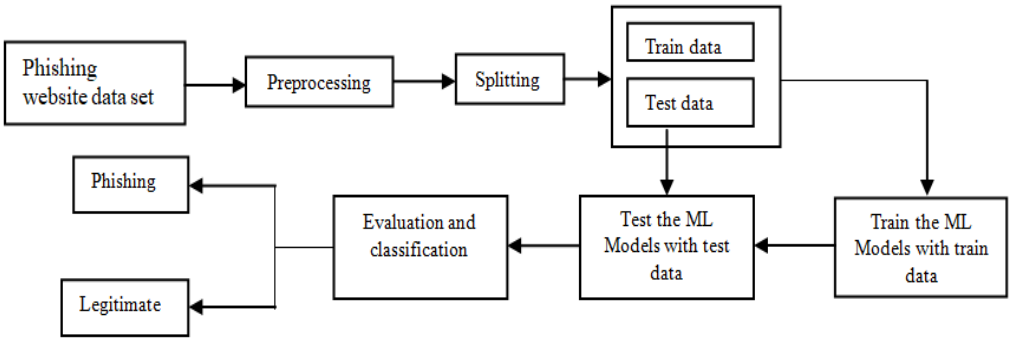


Figure 1 the proposed framework

**Proposed algorithm**

**Algorithm: Phishing Website Detection using Machine Learning (PWD-ML)**

Inputs:Phishing website data set, prediction models M

Output: Results as R

a. Start
b. Input dataset
c. Pre-processing
d. Extract features from training set()
e. Train the model m
f. End For
g. For each model m in M
h. Use model for testing
i. Evaluate
j. Display results
k. End For
l. End
m. Return R

As input, (PWD-ML) takes a specified dataset and generates a pipeline of multiple prediction models. It distributes the dataset 80:20 between training and testing data. A subsequent iterative procedure employs many ML models to detect phishing websites. Each model is assessed using distinct performance indicators. Each model generates its own confusionmatrix, which is then utilised to calculate precision, recall, F1-score, and accuracy. The algorithm gives both the findings of fraud detection and performance statistics.**3.2 Performance Evaluation Metrics**

In many ML-based situations, performance evaluation measures are derived from confusion matrices. According to [2, 3, 7, and 10], confusion matrix is extensively employed. There are two accurately predicted cases (TP and TN) and two incorrectly predicted cases (FP and FN).

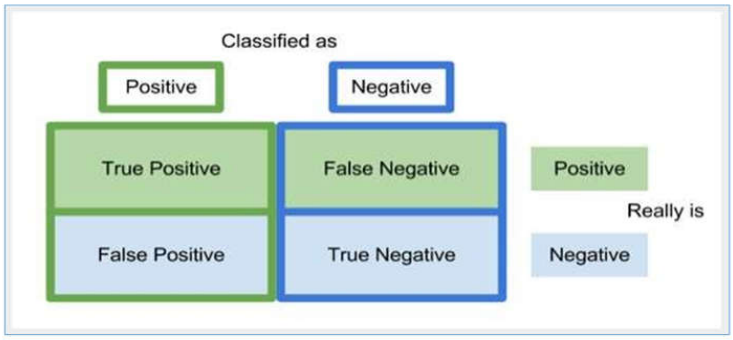


**Figure 3:** Confusion matrix model

As presented in Figure 3, different cases in the confusion matrix are used to arrive at the performance measures. The performance metrics are expressed in Eq. (1) to Eq. (4).

$$\text{Precision} = \frac{TP}{TP+FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (2)$$

$$\text{F1-measure} = 2 * \frac{(precision * recall)}{(precision+recall)} \quad (3)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

These metrics result a value between 0 and 1 reflecting lowest and highest possible performance. Higher value refers to better performance.

**Dataset description**

The collection of phishing URLs is compiled by the open-source programme PhishTank. This service provides an hourly-updated list of phishing URLs in several formats, such as CSVand JSON. To obtain the data, please visit https://www.phishtank.com/developer info.php. 5000 random phishing URLs are collected from this dataset to train machine learning models.

The valid URLs are obtained from the University of New Brunswick's open datasets at https://www.unb.ca/cic/datasets/url-2016.html. This dataset contains benign, spam, phishing, malicious, and defacement URLs. For this project, only the benign urldataset type is being considered. ML models are trained using 5000 random, valid URLs extracted from this dataset.

This data set was downloaded from the Machine Learning Repository, Center for Machine Learning and Intelligent Systems at the University of California, Irvine [11]. It includes content from 1353 URLs. There are 548 legal emails, 702 phishing emails, and 103 suspicious emails. Additionally, the dataset comprises nine features taken from each URL. The properties offer information such as the URL anchor, the popup window, the age of the domain, the length of the URL, the IP address, and online traffic, among others. Each feature value contains binary or ternary categorical values. Existence or absence of a feature within the URL determines the value assigned to that feature according to binary values. The value attributed to ternary features is determined by the feature's presence in a particular ratio. In the following paragraphs, we outline the characteristics that we employed in our research.

**Results and discussion**

This section explains the feature analysis, data analysis, and experimental outcomes performed on the chosen phishing detection data set. Characteristics analysis Several experiments have been conducted to examine the association between the individual characteristics and the outcome.

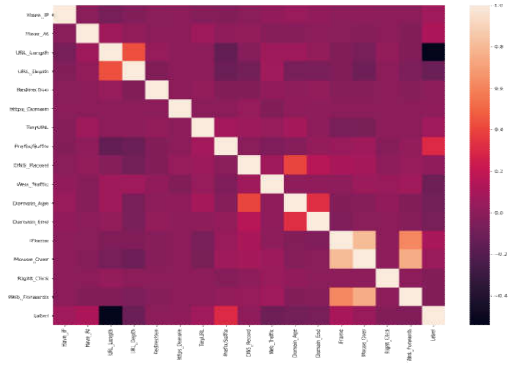

Fig 4 correlation of all features

**Random forest feature importance (RFE)**

RFE is also used to determine the significance of the feature set. Except for port, Iframe, rightclick, and on mouseover, the rating of all features was determined to be the same, that is, 1. These properties have the same range of values from -1 to 1 inclusive. Here, 1 indicates legitimacy, 0 indicates suspicion, and -1 indicates phishing. When a user submits information on a Web page, this data is sent to the authentication server from which the page was loaded. However, a phisher redirects the link to a different URL link than the correct URL. If a website requests personal information from users via a pop-up window, there is a potential that it is phishing. To fool users, phishers spoof HTTP protocols. If objects are loading from aURL other than the specified URL, the site is likely malicious. Phishing Web pages contain links that lead to domains other than the one specified in the URL.

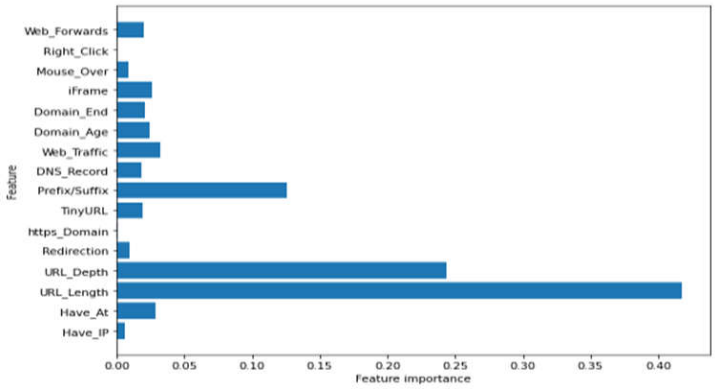


Fig 5 Random forest feature importance (RFE)

**Classification results and discussion**

Here, the experimental outcomes, parameter settings, projected model outcomes, and recommended features are addressed. To test the performance of the classifiers, all of the data set's features are normalised, along with the proposed features. After normalisation, significant features and multiple ML algorithms with tenfold cross validation settings are performed. Machine learning algorithms have been imported using Scikit-learn. The dataset is separated into a training set and a testing set in proportions of 50:50, 70:30, and 90:10, respectively. Each classifier is trained using the training set, while the testing set is utilised to evaluate their performance. Classifier performance has been evaluated by computing the accuracy score, false negative rate, and false positive rate of each classifier.

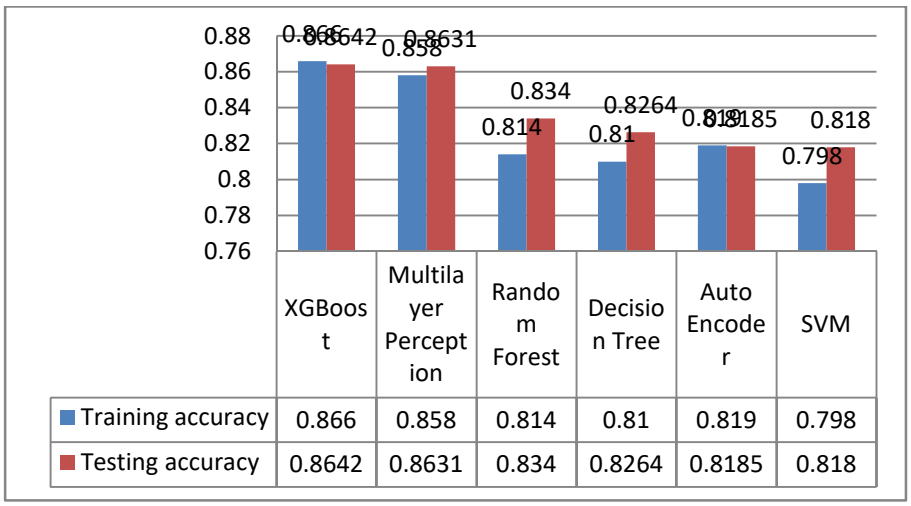


Fig 6 models performance comparison

As seen in Figure 6, ML is utilized to detect phishing websites. The accuracy of various models is compared to that of other employed models. When the proposed method is utilized, there is a large performance increase.

**Conclusion**

The goal of this study is to develop machine learning-based solutions for phishing website detection. Using the XG Boost algorithm, we were able to discover targets with an accuracy of 86.42 percent and the fewest false positives. Results also show that using more data as training data improves classifier performance. In this study, several algorithms and techniques for identifying phishing websites were provided by machine learning researchers. After reading the publications, we found that popular machine learning methods including XG Boost, SVM, Decision Tree, Random Forest, and auto encoder were used for the majority of the work. A different detection method, akin to Phish Score and Phish Checker, was suggested by certain authors. Utilized were the combinations of qualities in terms of recall, accuracy, etc. Successful experimental strategies for detecting phishing website URLs As the number of phishing websites continues to rise, some detection technologies may be added or replaced with new ones.

**References**

[1] Y. Sonmez, TurkerTuncer, HuseyinGokal&EnginAvci (2018). ‐Phishing web Sites Features Classification Based on Extreme Machine Learning‖. 6th International Symposium on Digital Forensic and Security (ISDFS).

[2] Kay, R (2004) Sidebar: The Origins of Phishing. [Online]. Available:http://www.computerworld.com/s/article/89097/Sidebar_The_Origi ns_of_Phishing. [3]Mohmoudkhonji, Youssef Iraqi and Andrew Jones(2013).‖Phishing detection: A Literature Survey‖. IEEE communications Systems and Tutorials, pp (99):1-31.

[4] Eduardo Benavides, Walter Fuertes, Sandra Sanchez & Manuel Sanchez(2019).‖Classification of Phishing Attack Solutions by Employing Deep Learning Techniques: A Systematic Literature Review'. Smart Innovation, Systems and Technologies, vol 152. Springer, Singapore.

[5] Alswailem, A., Alabdullah, B., Alrumayh, N., &Alsedrani, A. (2019)‖Detecting Phishing Websites Using Machine Learning‖.2nd International Conference on Computer Applications & Information Security (ICCAIS)

[6] Himani Thakur &Supreetkaur(2016).‖A Survey Paper On Phishing Detection‖. International Journal of Advanced Research in Computer Science(IJARCS). ISSN: 0976- 5697.

[7] Kathrine, G. J. W., Praise, P. M., Rose, A. A., &Kalaivani, E. C. (2019). ‐Variants of phishing attacks and their detection techniques‖. 3rd International Conference on Trends in Electronics and Informatics.

[8] R.Priya (2016), ‐An Ideal Approach for Detection of Phishing Attacks using Naive Bayes

Classifier‖. International Journal of Computer Trends and Technology(IJCTI). ISSN: 2231-2803.

[9] Arun Kulkarni, Leonard L.. "Phishing Websites Detection using Machine Learning", International Journal of Advanced Computer Science and Applications, 2019

[10] Aron Blam, Brad Wardman, Thamar Solorio and Gary Warner (2010),‖ Lexical feature based phishing URL detection using online learning‖, 3rd ACM Workshop on Security and Artificial Intelligence.

[11] Rami M. Mohammad, FadiThabtah& Lee McCluskey ‖Phishing Websites Features‖.(2014).

[12] Tyagi, I., Shad, J., Sharma, S., Gaur, S., & Kaur, G. (2018).‖A Novel Machine Learning Approach to Detect Phishing Websites‖. 5th International Conference on Signal Processing and Integrated Networks (SPIN).

[13] Singh, P., Maravi, Y. P. S., & Sharma, S. (2015). –Phishing websites detection through supervised learning networks‖. 2015 International Conference on Computing and Communications Technologies (ICCCT).

[14] Basnet, R., Mukkamala, S., & Sung, A. H. (n.d.). Detection of Phishing Attacks: A Machine Learning Approach. Studies in Fuzziness and Soft Computing, 373–383.

[15] Gyan Kamal and Monotosh Manna, Detection of Phishing Websites Using Naive Bayes Algorithm, Proceeding of International Journal of Recent Research and Review, Vol. XI, Issue 4 December 2018, ISSN 2277-8322.

[16] Baykara, M., &Gurel, Z. Z. mm(2018). Detection of phishing attacks. 2018 6th International Symposium on Digital Forensic and Security 355389(ISDFS).

[17] M. Kaytan and D. Hanbay ‖Effective classification of Phishing Webpages Based on New Rules by Using Extreme Machine Learning‖ Anatolian Journal of Computer Sciences, AJCS 17, pp: 15-36, ISSN: 2548- 1304, 2017.

[18] Xiang, G., Hong, J., Rose, C. P., &Cranor, L. (2011). CANTINA+A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites. ACM Transactions on Information and System Security, 14(2), 1–28.