International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT2390227

# A Review on Phishing Website Detection Using Machine Learning Approach

**Nikita Pawar[1], Dr. P. A. Tijare[2]**

[1]ME Student, Computer Science and Engineering, Sipna COET, Amravati, Maharashtra, India

[2]Professor, Computer Science and Engineering, Sipna COET, Amravati, Maharashtra, India

ARTICLEINFO

**Article History:**

Accepted: 01 April 2023
Published: 09 April 2023

**Publication Issue**
Volume 10, Issue 2
March-April-2023

**Page Number**
267-272

ABSTRACT

Phishing attacks are a rapidly growing threat in the cyber world. Every person is truly dependent on the internet. Everyone is doing online purchasing and online activities like online banking, online booking, and online billing on the internet. Phishing is a type of threat on a website and phishing is the illegal use of real information on the website information such as login id, password, information of credit card. Phishers use websites that visually and semantically resemble real websites. How can someone tell the difference between a website that is a phishing website and a real website? The identification of a website as a phishing website depends on several factors such as URL length, including the special letter '@', double slash redirect, and the existence of subdomains. Although the above factors are present on the website, no one can claim that the website is a phishing website, it can also be an original website. For solving this type of problem we can use the machine learning algorithm. The review creates phishing attack alerts, detects the attack, and motivate readers to use phishing prevention. With a large number of phishing emails or messages coming today, it is for businesses or individuals impossible to find them.

Keywords: Phishing website, Phishing URL Detection, Machine learning.

## I. INTRODUCTION

In 2020, due to the global pandemic people lives completely depended on technology. When digitization became relevant in this scenario, cybercriminals launched the online crime wave. Recent reports and investigations indicate an increase in security breaches that have cost victims large sums of money or disclosure of sensitive information. Phishing is a cybercrime that uses both social engineering and technical deception to steal personal information or financial account credentials from victims. Phishing involves attackers spoofing trusted websites and redirecting people to those websites, where they are tricked into sharing usernames, passwords, bank or credit card information, and other sensitive credentials. These phishing URLs can be sent to consumers via email or SMS. According to the FBI Crime Report 2020, phishing was the most common type of cyberattack in 2020, with nearly phishing incidents doubling from 114,702 in 2019 to 241,342 in 2020. Data breaches in 2020 involved phishing. The

number of phishing attacks observed by the Anti Phishing Work Group (APWG) increased in 2020, doubling from this year. Phishing attacks on SaaS and webmail sites decreased and attacks on e-commerce sites increased, while attacks on media companies fell slightly from 12.6% to 11.8%. In light of the current pandemic, has experienced a series of phishing attacks in which leverages it is global focus on Covid-19.

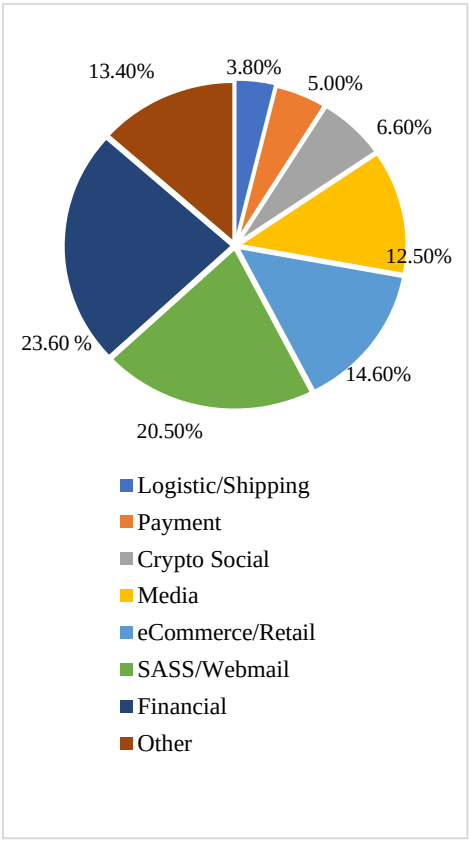


Fig. 1. Most Phishing Sites

According to WHO, many hackers cybercriminals send fraudulent emails and According to the author [4] various classifiers such as Logistic Regression, Naive Bayes classifiers, Random Forest, Decision Tree, and K-Nearest Neighbor formed based on features extracted from the lexical structure of URL.

The data set contained an equal number of WhatsApp messages to people. These attacks come in the form of fake job listings, fabricated messages from health organizations, Covid vaccine phishing, and branding impersonation [1]. Therefore it becomes more and more necessary to secure this site. Protocols and regulations protect the connection between client and server but remain vulnerable to a malicious attack. The term "malicious" is an umbrella term for attack types that include phishing, spam and malware and more [8].

According to the Anti-Phishing Working Group (APWG) report phishing activity trends the total number of phishing sites observed in a phishing and legitimate URLs and was created to address data bias, learning bias, variance and overfitting issues in a 7:3 ratio for the training and test.

The author [5] proposed a machine learning- based phishing detection system that uses 8 different algorithms on three different sets of data. Algorithms used Logistic Regression (LR), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Naive Bayes (NB), XGBoost, Random Forest (RF) and Artificial Neural Network (ANS). It has been observed that models with LR, SVM and NB have a low accuracy rate. In terms of training time, the algorithms NB, DT, LR, and

ANN performed better. They concluded that the RF algorithm or ANN algorithm could be used because of the shorter training time and high accuracy rate.

Table 1 : Accuracy Rate of Different Machine Learning Algorithms

| Paper | Algorithm | AccuracyRate |
|---|---|---|
| Iman Akouret al. [3] | SVM | 96.3% |
| | KNN | 94.0% |
| | LR | 93.5% |
| | NB | 89.7% |
| Jitendra Kumar e tal. [4] | LR | 97.7% |
| | RF | 98.3% |
| | NB | 97.1% |
| | DT | 98.2% |

| | | |
|---|---|---|
| | KNN | 97.9% |
| Mehmet Korkmaz et al. [5] | LR | Dataset 1- 91.3%<br>Dataset 2- 75.6%<br>Dataset 3- 78.2% |
| | KNN | Dataset 1- 91.4%<br>Dataset 2- 81.4%<br>Dataset 3- 81.1% |
| | SVM | Dataset 1- 87.3%<br>Dataset 2- 70.0%<br>Dataset 3- 76.7% |
| | DT | Dataset 1- 92.5%<br>Dataset 2- 81.6% |
| | | Dataset 3- |
| | | 81.6% |
| | NB | Dataset 1- |
| | | 88.3% |
| | | Dataset 2- |
| | | 70.0% |
| | | Dataset 3- |
| | | 67.0% |
| | XGBoost | Dataset 1- |
| | | 92.9% |
| | | Dataset 2- |
| | | 83.6% |
| | | Dataset 3- |
| | | 83.2% |
| | RF | Dataset 1- |
| | | 94.5% |
| | | Dataset 2- |
| | | 90.5% |
| | | Dataset 3- |
| | | 91.2% |
| | ANN | Dataset 1- |
| | | 94.3% |
| | | Dataset 2- |
| | | 88.2% |
| | | Dataset 3- |
| | | 88.8% |
| Mohammad Nazmul et al. [6] | DT | 91.9% |
| | RF | 96.9% |
| Ilker Kara et al. [7] | KNN | 96.5% |
| | SVM | 96.6% |
| | DT | 96.2% |
| | RF | 96.0% |
| | LR | 94.5% |
| | LDA | 94.6% |

Table 1, shows the accuracy rate of various algorithms in machine learning.

The author [6] proposed a system for detecting phishing attacks using a random forest and decision tree. A Kaggle dataset with 32, features was used with feature selection algorithms such as principal component analysis (PCA). Selecting a feature reduces the redundancy of irrelevant or unnecessary data in the data set. The proposed model used the REF, Bump-F, IG, and GR algorithms to select features prior to application PCA.

The author [7] used the machine learning technique for detection of phishing URL by using these methods K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Linear Discriminant

Analysis (LDA). The methods were used to compare the acquired training datasets. Because phishing sites and legitimate sites are different.

## II. COMPARATIVE ANALYSIS

Comparison of various machine learning algorithms.

3.1 Support Vector Machine:

Support Vector Machine (SVM) is a very simplest algorithm for implement so we can work with huge number of input variable and can create effective solution using basic tricks. Hence it is inappropriate for large database and if there crash it cannot perform well.

3.2 Decision Tree:

A Decision Tree (DT) is an algorithm that divides a data set of into subsections to form a tree. In the tree created by each node is associated with a feature and each leaf with a class. This easy to explain algorithm consisting of a few hyperparameters, may not be efficient with multiple classes or small data sets.

3.3 K-Nearest Neighbor:

K-Nearest Neighbor (KNN) is a fast and structured algorithm that generates approximately based on the distances of k neighbors. Although counting this distance with large amounts of data means a lot of memory usage.

3.4 Naive Bayes:

Naive Bayes (NB) is a classification algorithm this goes to Bayes theorem. It is simple for of implementation and reduced training time. However, this is not preferred, for example for reasons such as a lower estimate with fewer data.

3.5 Logistic Regression:

The Logistic Regression (LR) algorithm can produce good predictions only when the based on the variable is collected into two classes. It is an advantage in this factors such as feature repetition and also outliers negatively impact the prediction.

3.6 Random Forest:

Random Forest (RF) is an algorithm that works with a set of learning technique to create a huge number of trees in the data set and be less prone to noise. However, the trains a lot and needs processor power and memory.

3.7 Artificial Neural Network:

Artificial Neural Network (ANN) is an algorithm that works with at minimal three layers of a structure. During the training of the model, it can recognize connections between aspects and generalize well. It can require lots of memory.

3.8 XGBoost:

XGBoost is an algorithm is gradient-based decision trees that prioritize speed and efficiency. It is purpose is to decrease the number of faults in the previous tree by creating a new tree each time. Although this process can take a more time.

3.9 Linear Discriminant Analysis:

Linear Discriminant Analysis (LDA) is a dimensionality easing technique used in machine learning classification problems. LDA also breaks in a few cases where the mean of distributions is common.

According to these data consistency of KNN, LR, RF, DT algorithms are nearby same as compare to the SVM, NB, XGBoost, LDA, ANM algorithms.

## III. PROBLEM STATEMENT

Phishing sites are fake sites that can be constructed and impersonated as legitimate sites to trick other people into stealing their vital personal information such as bank account information, social security numbers, and passwords. It will cause an information security breach by stealing sensitive data, resulting in

financial losses for the victim. In short, this is an internet scam or crime of the highest order.

Therefore, the assessment or detection of phishing sites requires an intelligent model to detect and detect suspicious characteristics associated with phishing sites. The main issue addressed in this study is how to strengthen user authentication on the website. The study examines the potential applications of three classification models developed by the team in detecting phishing sites. Specifically, the goal here is to develop an aggregate model that will be used to predict whether a website is phishing or legitimate and to measure the accuracy of the site detection of phishing to improve [1].

## IV. CONCLUSION

Phishing detection is currently of great interest to researchers due to its importance to privacy and security. There are several methods to detect phishing by classifying websites us in trained machine learning models. The URL- based analysis increases recognition speed. Also, by using feature selection algorithms and dimension reduction techniques, we can reduce the number of features and remove irrelevant data. There are many machine learning algorithms that perform classification with good performance measurements. In this paper, we conducted a study on phishing detection and the phishing detection process using machine learning algorithms. This serves as a adviser for new researchers to know the process and develop more accurate phishing detection systems.

## V. REFERENCES

[1]. Arathi Krishna V, Anusree A, Blessy Jose, Karthika Anilkumar, Ojus Thomas Lee, "Phishing Detection using Machine Learning based URL Analysis: A Survey", International Journal of Engineering Research & Technology (IJERT), Volume 9, Issue 13, pp.156-160, 2021.

[2]. Ammar Odeh, Ismail Keshta, Eman Abdelfattah, "Machine Learning Techniques for Detection of Website Phishing: A Review for Promises and Challenges", 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC), DOI:10.1109/CCWC51732.2021.937599, 2021.

[3]. Iman Akour, Noha Alnazzawi, Ahmad Aburayya, Raghad Alfaisal, Said A. Salloum, "Using Classical Machine Learning For Phishing Websites Detection From URLS", Journal of Management Information and Decision Sciences, Volume 24, Special Issue 6, 2021.

[4]. Jitendra Kumar, Balaji Rajendran, A. Santhanavijayan, B. Janet, Bindhumadhava BS, "Phishing Website Classification and Detection Using Machine Learning", 2020 International Conference on Computer Communication and Informatics (ICCCI -2020), 2020.

[5]. Korkmaz, Ozgur Koray Sahingoz, Banu Diri, "Detection of Phishing Websites by Using Machine Learning- Based URL Analysis", 11nth International Conference on Computing Communication and Networking Technologies (ICCCNT), 2020.

[6]. Mohammad Nazmul Alam, Dhiman Sarma et al., "Phishing attacks detection using machine learning approach," 3rd International Conference on Smart Systems and Inventive Technology (ICSSIT), 2020.

[7]. Ilker Kara, Murathan Ok,Ahmet Ozaday, "Characteristics of Understanding URLs and Domain Names Features: The Detection of Phishing Websites With Machine Learning Methods",IEEE Access, Volume 10, 2022, DOI:10.1109/ACCESS.2022.3223111, pp.124420-124428, 2022.

[8]. Malak Aljabri, Hanas S. Altamimi, Shahd A. Albelali, Maimunah Al-Harbi, Haya T. Alhuraib, Najd K. Alotaibi, Amal A. Alahmadi,Fahd Alhaidari, Rami Mustafa

[9]. A. Mohammad, Khaled Salah, “Detecting Malicious URLs Using Machine Learning Techniques: Review and Research Directions”, IEEE Access, Volume 10, 2022, DOI:10.1109/ACCESS.2022.3222307, pp. 121395-121417, 2022.

**Cite this article as :**

Nikita Pawar, Dr. P. A. Tijare, "A Review on Phishing Website Detection Using Machine Learning Approach", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 9, Issue 2, pp.267-272, March-April-2023. Available at doi : https://doi.org/10.32628/CSEIT2390227
Journal URL : https://ijsrcseit.com/CSEIT2390227