International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 6, Nov-Dec-2024, ISSN (Online): 2395-566X

Detection of Phishing Websites Using Machine Learning

Manish Gujral, Harsh Kumar, Annu Sharma, Dr.Monika
Department of Computer Science Engineering
Manav Rachna International Institute of Research and Studies

Abstract- Phishing is a category of cyberattack that includes the theft of credit card numbers, passwords, and other private data. We have employed machine learning algorithms to identify phishing websites in order to prevent phishing fraud. The availability of several services, including social networking, software downloads, online banking, entertainment, and education, has sped up the development of the Web in recent years. Consequently, enormous volumes of data are downloaded and uploaded to the Internet on a regular basis. Attackers can now obtain private information, including social security numbers, account numbers, passwords, and usernames, as well as financial information. This is one of the most important problems with web security and is referred to as a "phishing" attack on the internet. To identify these malicious websites, we employ a variety of machine learning methods, including KNN, Naive Bayes, Gradient Boosting, and Decision Trees. The study is broken down into the following sections. The introduction outlines the tools, methods, and concentrated zones that are employed. The process of gathering the data needed to proceed is described in depth in the preliminary section. Subsequently, the paper highlights the thorough examination of the information sources.

Index Terms- Phishing Detection, Phishing, Security threats, Machine learning, Datasets, Algorithms, Cyber Security.

I. INTRODUCTION

Phishing is a cybercrime that attempts to deceive gullible people into disclosing private information that is both lucrative and sensitive. The attacker then exploits these facts for nefarious objectives, such as identity theft, and can include login credentials, financial account information, usernames, passwords, personal addresses, and social media connections. Phishing attacks are very easy to deploy and don't require much in the way of technological complexity. Still, they work pretty well most of the time. It can be very difficult for people to recognize phishing sites because attackers craft well-designed phishing websites that mimic the appearance and feel of genuine websites. Additionally, as shown in [1], attackers have improved their evasion and tactic strategies over time to avoid being discovered. These days, one of the key areas of computer science is cybersecurity, which deals with preventing threats to and safeguarding user data from attacks. It ensures that users won't be tricked by these attacks and end up as victims of online crimes [2]. There are numerous direct and indirect effects of phishing attacks. They have an impact on the victims of phishing, whose identities and accounts may be compromised, potentially raising suspicions about online services and resulting in financial theft. The businesses and organizations that are being impersonated are also impacted by these attacks, as their brands may be misused and result in data breaches, monetary losses, and reputational harm. The need for a thorough insight study in order to examine phishing website detection methods is what spurred this review. There was no systematic format for a collaborative work that contains the methods, datasets, and algorithms used in phishing website detection. A thorough analysis and study of this field were required.

This study aims to assess and make a contribution to the following:

- Learn the most effective methods for identifying phishing websites so that the security manager can quickly choose the best anti-phishing strategy for their security systems.
- The methods, datasets, and algorithms employed by the researchers in the field should be the main emphasis of a strong review paper.

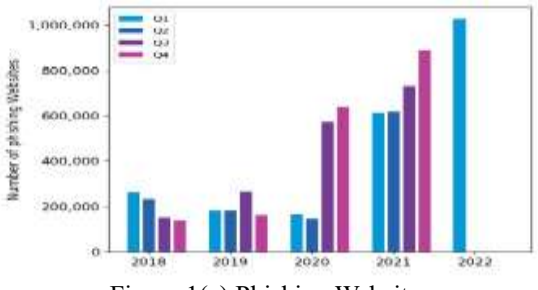

Figure 1(a) Phishing Websites

The Anti-Phishing Working Group (APWG) reported in its Phishing Activity Trends Report [3] that there were more than a million phishing websites detected in the first quarter of 2022. This surge has been especially noticeable since the start

of the COVID-19 pandemic, as Figure 1(a) illustrates. These consequences can also be seen in the distinct brands that phishing campaigns target, which have become more numerous since the third quarter of 2020 (see Fig. 1(b)).

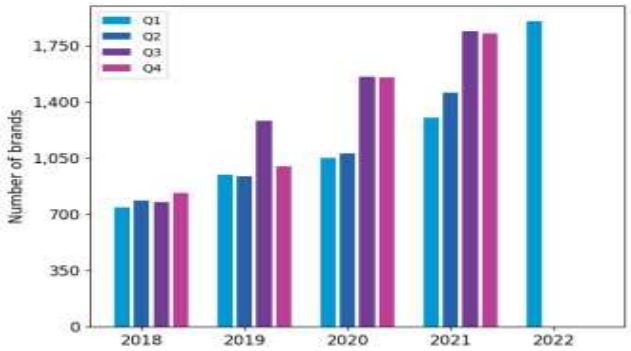

Figure 1(b) Brands

II. LITERATURE SURVEY

After careful examination of the writings of numerous authors, the following facts are finally revealed in this section of the survey.

[4] A Machine Learning Approach for Detection of Phished Websites Using Neural Networks by Charmi J. Chandan, Hiral P. Chheda, Disha M. Gosar, Hetal R. Shah.:Through web spoofing, people are lured to interact with false websites rather than the real ones. This attack's main objective is to steal users' private information. Shadow' website that mimics the original website is created by the attacker. Through this trick, the attacker can see and change any data that the victim submits. This research offers a method for identifying phishing websites by looking at the Uniform Resource Locators (URLs) on web pages. The suggested method makes a distinction between legitimate and fraudulent websites by looking up the Uniform Resources Locators (URLs) of suspicious online pages. Specific features of URLs are analyzed in order to identify phishing web pages.

[5] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. "Machine Learning-Based Phishing Detection from URLs," Expert Systems with Applications, 117:345-357, January 2019: The used dataset was created by the user. where authentic URLs come from Yandex Search API and phishing websites are owned by PhishTank. Finding words that resembled brand names, as well as keywords and words made up of random characters, was the major goal. Numerous algorithms are used for classification, including Naive Bayes, Random Forest, kNN(n=3), Adaboost, K-star, SMO, and Decision Tree. Certain feature extraction types are also used, including NLP-based features, Word Vectors, and Hybrid. The system maintains a high level of accuracy during the test.

[9] Dipayan Sinha, Dr. Minal Moharir, Prof. Anitha Sandeep, "Phishing Website URL Detection using Machine Learning," International Journal of Advanced Science and Technology, 29(3):24952504, 2020.: Machine learning techniques like Logistic Regression, Decision tree, Random Forest, Adaboost, Gradient Boosting, Gaussian NB, and Fuzzy pattern tree classifier are used to detect phishing websites. Both authentic and phishing websites are used in data collection. There are two steps in extracting useful features: IP addresses, the @ symbol, dashes, lengthy URLs, the presence of unusual numbers, dot counts, subdomains in URLs, and other elements are all part of URL-based computing. Domain-based factors include the website's Page Rank, its age, and its validity. The dataset is divided 80:20 into training and testing sets. The Random Forest algorithm has the highest F1 score of 95% and displays 96% precision and recall rates.

III. TECHNIQUES

Several popular machine learning classification methods, such as the following, have shown to be effective in identifying phishing domains:

1. Decision Tree

Through the use of a tree-like graph or modeling of alternatives and their potential consequences, including utility, resource costs, and likelihood of outcomes, a decision tree assists people in making better decisions. To illustrate an algorithm composed entirely of conditional control statements, this is just one method among many [12]. Big datasets can be analyzed by using decision trees to examine the underlying relationships. The purpose of the decision tree is to watch a process. By doing so, researchers can use its characteristics to classify the process, as demonstrated in Figure 1, which depicts a training algorithm that builds the decision tree's structure. Depending on how it is constructed, a decision tree can be used to evaluate additional samples with varying degrees of success.

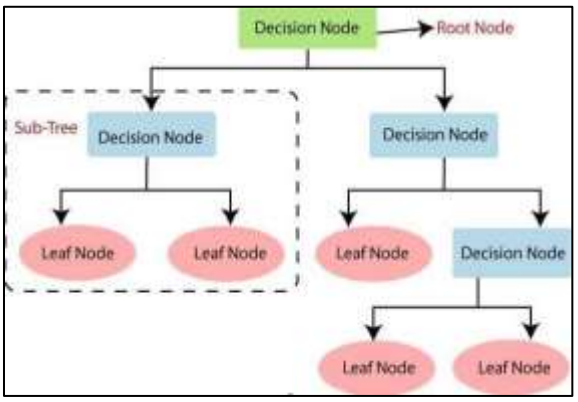

Figure 2. Decision tree algorithm

International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 6, Nov-Dec-2024, ISSN (Online): 2395-566X

2. Random Forest

In machine learning and predictive modeling, a random forest is a collection of supervised learning algorithms for regression and classification [13]. Academics are interested in the random forest because of its quick and precise categorization. In order to select the optimal output—the mode of the classes, or the value that appears most frequently in the decision tree results—it compiles the outcomes and predictions of multiple decision trees. The training set and the test set are the two portions of the dataset that Random Forest divides into. Then, it chooses a number of samples at random from the training set.

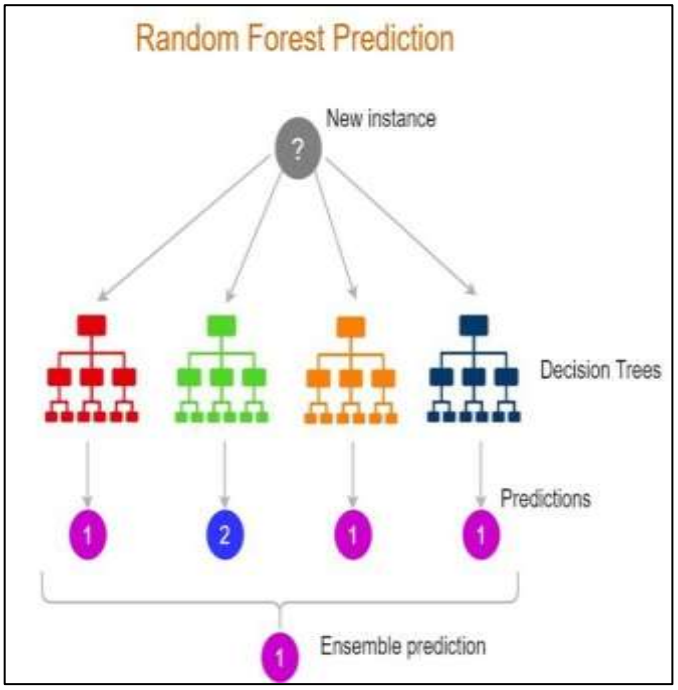

Figure 3. Random Forest Prediction

3. Support Vector Machine

SVM is a supervised learning technique used for regression and pattern recognition that is based on statistical learning theory. While statistical learning theory can identify the conditions under which a particular, simple algorithm can be learned, realworld applications typically call for more complex tools and algorithms (like neural networks), which are far more challenging to theoretically analyze.

The intersection of learning theory and practice is found in SVMs. They produce models that are both sufficiently simple to be studied mathematically and sufficiently complex (such as a large class of neural networks, for example). An SVM is a linear algorithm in a high-dimensional space, which explains why [8]. Figure 3 illustrates how SVM predicts labels by creating a decision boundary, or hyperplane, with a minimum of one between two given classes.

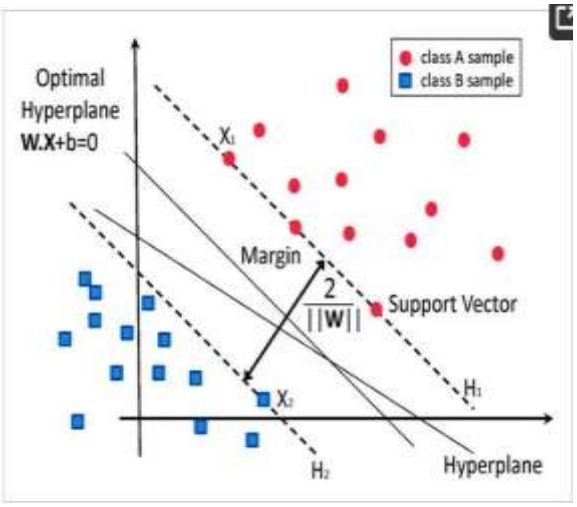

Figure 4. Support Vector Machine

4. Ensemble Classification Techniques

Creating a fair model from a dataset is one of the primary objectives of algorithms for machine learning. The process of creating models from data is called learning, or training, and the model that has been learned is known as a hypothesis or learner. Using ensemble methods, new data points are classified into groups based on the predictions made by a set of classifiers that are created by learning algorithms [14].

When compared to the individual classifiers that comprise them, ensembles exhibit far higher accuracy. To train multiple hypotheses to solve a problem, ensemble methods—also referred to as committee-based learning or learning multiple classifier systems—are employed. Several decision trees are used in random forest trees, a popular type of ensemble modeling, to predict results. A general ensemble architecture is shown in Figure 5.

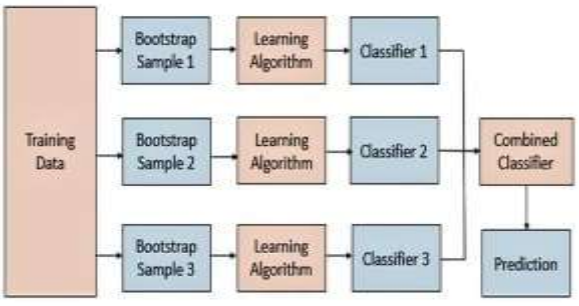

Figure 5. Ensemble Learning

IV. METHODOLOGY

A phishing website is a kind of social engineering wherein authentic websites and uniform resource locators (URLs) are mimicked. The most frequent method used by phishing

attacks is the Uniform Resource Locator (URL). The sub-domains of the URL are entirely under the phisher's control. Because the URL includes file components and directories, the phisher can change it.

Process

1 Data Collection

The open-source program Phish Tank was used to compile the list of phishing URLs. This website offers an hourly updated list of phishing URLs in several formats, such as csv, json, and others.

2. Data Cleaning & Data PreProcessing

To clean up the data, add missing numbers, smooth out crooked data, find and remove outliers, and fix anomalies.

The process of cleaning unstructured raw data to create a tidy, well-organized dataset that can be utilized for additional research is known as data preprocessing. Data preprocessing is a cleaning procedure that creates a clean, well-organized dataset that can be utilized for additional research out of unstructured raw data.

3. Address & Domain Based Checking

The categories that were taken from address-based:

- Domain of the URL
- "@" Symbol in URL
- IP Address in the URL
- Length of URL
- Depth of URL
- DNS Record
- Web Traffic
- Age of Domain
- End Period of Domain

4. Machine Learning Models

Decision Tree Classifier

Decision trees are a popular model for applications involving regression and classification. In essence, they are taught a hierarchy of if/else questions that result in a selection. The process of learning a decision tree involves committing the series of if/else questions to memory that, when answered correctly, will take the least amount of time. In order to construct a tree, the algorithm iterates through each possible test to find the most informative one regarding the target variable.

Random Forest Classifier

One of the most popular machine learning techniques for regression and classification is the random forest classifier. All that is a random forest is an arrangement of decision trees, each varying slightly from the others. The idea behind random forests is that, even though each tree might predict fairly well, on some data it will almost definitely overfit. They usually function well with little parameter adjustment, are very powerful, and don't require scalable data.

5. Evaluation

Decision Tree Classifier: To construct a tree, the algorithm iterates through all possible tests, selecting the one that provides the most information about the target variable. Here, we forecast the model's accuracy using data collected from both test and trained samples. The accuracy of the test and training datasets, respectively, was found to be 82.6% and 81%. The decision tree classifier algorithm in action is shown below. The model is fitted in the tree after a number of parameters are set. To assess the model's accuracy, the samples are split into X and Y train and X and Y test groups.

Random Forest Classifier: By averaging the results of multiple trees that all function well and overfit in different ways, we can reduce the amount of overfitting. The first step in building a random forest model is figuring out how many trees to build. They usually function well with little parameter adjustment, are very powerful, and don't require scalable data.

Flowchart

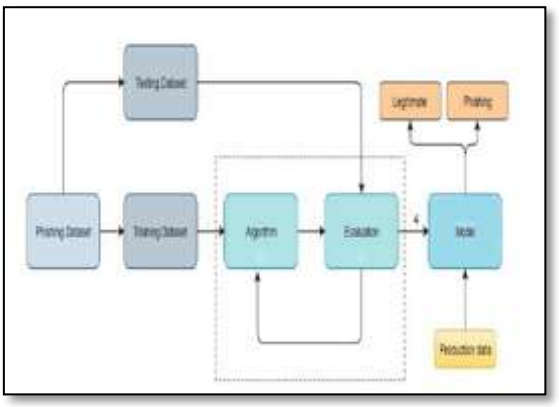

Figure 6: Flowchart of Proposed Model

V. CONCLUSION

In this study, we demonstrated the effectiveness of machine learning techniques in detecting phishing websites by analyzing various features such as URL structure, domain age, SSL certification, and content-based factors. By applying algorithms like [insert algorithms used: e.g., Random Forest, SVM, etc.], we achieved a high detection accuracy, proving that machine learning models can effectively differentiate between legitimate and phishing sites.

The results indicate that automated detection using machine learning reduces human error and speeds up identification, making it a crucial tool in enhancing online security. However, continual updates to the model are necessary to keep up with evolving phishing tactics.

Future work could focus on integrating deep learning and real-time detection systems to further enhance performance and adaptability.

This approach represents a significant step forward in combating phishing threats and can be implemented in various cybersecurity frameworks to protect users and organizations from online fraud.

REFERENCES

1. P. Zhang, A. Oest, H. Cho, Z. Sun, R. Johnson, B. Wardman, S. Sarker, A. Kapravelos, T. Bao, R. Wang, Y. Shoshitaishvili, A. Doupé, and G.-J. Ahn, ‘‘CrawlPhish: Large-scale analysis of client-side cloaking techniques in phishing,’’ in Proc. IEEE Symp. Secur. Privacy (SP), May 2021, pp. 1109–1124.
2. Nikhita Reddy, G.J. Ugander Reddy, “A Study Of Cyber Security Challenges And Its Emerging Trends On Latest Technologies,” International Journal of Engineering and Technology, vol. 4, no.1, January 2014.
3. Anti-Phishing Working Group—APWG. (2022). Phishing Activity Trends Report-1Q. [Online]. Available: https://docs.apwg.org/reports/apwg_trends_report_q1_2022.pdf.
4. A Machine Learning Approach for Detection of Phished Websites Using Neural Networks by Charmi J. Chandan, Hiral P. Chheda, Disha M. Gosar, Hetal R. Shah.
5. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. “Machine Learning-Based Phishing Detection from URLs,” Expert Systems with Applications, vol. 117, pp. 345-357, January 2019.
6. Vayansky, I. and Kumar, S., “Phishing – challenges and solutions.”, Computer Fraud & Security, vol 2018, no. 1, pp. 15-20, January 2018.
7. Nikhita Reddy, G.J. Ugander Reddy, “A Study Of Cyber Security Challenges And Its Emerging Trends On Latest Technologies,” International Journal of Engineering and Technology, vol. 4, no.1, January 2014.
8. Vahid Shahrivari, Mohammad Mahdi Darabi, Mohammad Izadi, “Phishing Detection Using Machine Learning Techniques,” unpublished.
9. Dipayan Sinha, Dr. Minal Moharir, Prof. Anitha Sandeep, “Phishing Website URL Detection using Machine Learning,” International Journal of Advanced Science and Technology, vol. 29, no. 3, pp. 2495-2504, 2020.
10. P. L. Indrasiri, M. N. Halgamuge, and A. Mohammad, ‘‘Robust ensemble machine learning model for filtering phishing URLs: Expandable random gradient stacked voting classifier (ERG-SVC),’’ IEEE Access, vol. 9, pp. 150142–150161, 2021.
11. R. Kiruthiga, D. Akila, “Phishing Websites Detection Using Machine Learning,” International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2S11, pp. 11114, September 2019.
12. Zhu, E.; Ju, Y.; Chen, Z.; Liu, F.; Fang, X. DTOF-ANN: An Artificial Neural Network Phishing Detection Model Based on Decision Tree and Optimal Features. Appl. Soft Comput. 2020, 95, 106505. [Google Scholar] [CrossRef]
13. Breiman, L. Random Forests. Mach. Learn. 2001, 45, 5–32. [Google Scholar] [CrossRef][Green Version]
14. Gomes, H.M.; Barddal, J.P.; Enembreck, F.; Bifet, A. A Survey on Ensemble Learning for Data Stream Classification. ACM Comput. Surv. CSUR 2017, 50, 1–36.
15. R. Kiruthiga, D. Akila, “Phishing Websites Detection Using Machine Learning,” International Journal of Recent Technology and Engineering (IJRTE), vol. 8, no. 2S11, pp. 11114, September 2019.