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 Website Detection using Classification Algorithms

Isha Nalawade[1], Sanjeevan Bapat[2]

[1]Final Year Student, Information Technology Department, Thadomal Shahani Engineering College, Bandra(W), Mumbai - 400050
[2]Final Year Student, Information Technology Department, Thadomal Shahani Engineering College, Bandra(W), Mumbai - 400050

Abstract - *Phishing is a sort of social engineering in which an attacker sends a fake communication in order to fool a person into disclosing sensitive information to the attacker or to install harmful software, such as ransomware, on the victim's infrastructure. It is critical to correctly classify phishing websites in order to detect and prevent phishing assaults. If a phishing assault has already happened, the classification of phishing websites can be used to establish recovery methods. Phishing website classification is a well-known engineering research topic. Machine Learning is commonly utilized in the identification of phishing websites because of its benefit of discovering essential traits from a dataset of multiple websites. The goal of this study is to address the problem of phishing website classification utilizing various classifiers, and ensemble learning. Ensemble learning approaches are used to enhance a classifier's performance. Extensive tests were conducted on the well-studied open access data collection "Phishing Testing Dataset" in this paper. Measures like f1-score, accuracy, recall and precision have been employed to evaluate the various models. The suggested approach has a remarkable accuracy of 97% in classifying phishing websites, according to experimental data. The proposed model would be viable in helping cyber-security experts and also the general public recognizes phishing websites accurately.*

Key Words: **Phishing, Websites, Detect, Machine-Learning, Classification, Ensemble**

1. INTRODUCTION

1.1 HISTORY:

"Love Bug", which is also known as, the first phishing email, is considered to be the reason many people learned about phishing. On May 4, 2000, several people fell victim to the Love Bug. Beginning in the Philippines, mailboxes throughout the world were flooded with the message "ILOVEYOU." "Kindly check the attached LOVELETTER coming from me," the message body stated. Those who couldn't stop themselves from discovering their hidden crush downloaded what they believed was a harmless .txt file, just to release a worm that harmed their local computer. The worm overwrote picture files and transmitted a clone of its own to all of the user's Outlook contacts. The history of phishing reveals that phishers' strategies have stayed pretty similar. The growth of social media has been a significant change. Social networking sites are a virtual treasure trove of personal data that fraudsters may and do use to customize emails to individual recipients, a technique known as spear phishing. The high stakes and the low resources needed to carry out an assault have made spear phishing the preferred method for thieves pursuing access to confidential information housed in the systems of major organizations and enterprises.

1.2 DEFINITION-

Phishing is the practice of delivering deceptive messages or emails that seem to originate from a trustworthy source. Emails are the most common method of carrying out phishing attacks. The main purpose of the attacker is to embezzle sensitive details such as credit card information or login credentials or to install any kind of malware on the personal computer of the victim. Everyone should be aware of phishing attacks in order to stay safe from such attacks.

A phishing domain is a website that looks and sounds similar to an official website. They are created in order to deceive someone into thinking it is authentic.

1.3 CHARACTERISTICS OF PHISHING WEBSITES:

The URL (Uniform Resource Locator) is utilized to locate any resource on the World Wide Web (or WWW).

A hostname is made up of a domain name and a sub-domain name. The phishing attacker has complete command over the name of the subdomain and route. The attacker has the ability to register any domain name that has not previously been registered. It can only be modified once. The unique element of the website makes it difficult for security guards to detect phishing websites. Once the fake domains have been identified, it is simple to thwart

the users from accessing them. The following diagram portrays the critical parts of the structure of a URL.

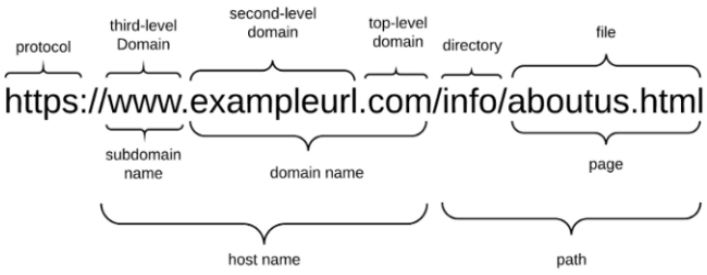

Fig -1: Characteristics of Phishing Website

Other methods of Phishing attacks:

Cybersquatting (also known as domain squatting) is the practice of registering, trading in, or utilizing a website address with the goal of profiting on the goodwill of someone else's brand. The cybersquatter may offer to sell the domain at an exorbitant value to an individual or corporation that owns a trademark included within the name, or he or she may utilize it for fraudulent reasons such as phishing.

Typosquatting, also known as URL hijacking, depends on mistakes made by Internet users when entering a website link into an internet browser or on typographical mistakes that are difficult to catch when reading. URLs established using Typosquatting appear to be reputable domains. A user may input a wrong website URL or click a link that appears to be from a trustworthy domain, leading them to an alternate website hosted by a phisher.

1.4 MACHINE LEARNING IN PHISHING DETECTION

Machine Learning (ML) is a form of AI (Artificial Intelligence) methodology, which uses numerous statistical, optimization, and probabilistic strategies to increase performance based on past experiences and fresh data. A wide range of ML approaches has been used widely to identify any phishing websites and prevent their consequences.

Several ML methods are quite effective in the early identification of phishing websites. Using ML algorithms, it is simple to discern between legitimate and counterfeit websites. Accurate classification will help people to detect phishing websites before their local system is attacked and the data is compromised. Additionally, classification is a complicated supervised optimization problem. For categorizing phishing domains, several classification approaches such as SVM, KNN, and Naive Bayes are employed.

2. RELATED WORK

This section addresses several pieces of literature on phishing website categorization. For the identification of phishing websites, several machine-learning techniques have been used.

In [1], the authors have applied two algorithms to determine if a website (URL) is counterfeit or genuine. The proposed solution trained the model using Random Forest Classifier and Decision Tree Algorithm to classify the websites. The model recorded an accuracy of 97% for the Random Forest Algorithm.

The authors, in [2], recommended utilizing Machine Learning approaches to detect phishing sites. The authors have employed the Random Forest Algorithm to classify phishing websites.

In [3], Rishikesh Mahajan and Irfan Siddavatam applied three Machine algorithms SVM, Decision Tree, and Random Forest algorithms on the phishing websites dataset. Among three machine learning algorithms, the Random Forest classifier recorded an accuracy of 97.14%.

[4] provided a detailed survey about the usage of ML methodologies for the identification and deterrence of phishing websites. The authors present a phishing website detection model involving the selection of optimal features and neural networks. The fuzzy rough set hypothesis has been employed to identify the most impactful features from a set of databases. The four classification models used to classify the websites are the Kernel Support Vector Machine (SVM), the Decision Tree Classifier, the K-Nearest Neighbor (K-NN), and the Random Forest Classifier.

Using the Random Forest Classifier, the authors of [5] present a methodology to determine phishing websites by employing a URL identification technique. The dataset is gathered from Phishtank. The suggested technique is divided into three phases parsing, followed by heuristic data classification, and lastly performance analysis. Only 8 of the 31 characteristics are evaluated for parsing. The random forest approach achieved a 95% accuracy level.

The c4.5 decision tree approach has been utilized in the paper [6] to demonstrate an efficient way to detect phishing sites. This approach collects webpage characteristics and generates the heuristic values. These heuristic values were sent into the c4.5-decision tree method, which determined if or not the target website was a genuine website or a phishing website. The dataset was gathered via sources like PhishTank and Google. This procedure was divided into two stages: pre-processing and detection. Here, the features are retrieved on the basis

of the rules defined in the first stage i.e. the pre-processing stage, and the attributes along with their values were input.

3. MATERIALS AND METHODS

Phishing websites are identified in this study utilizing classifiers such as Random Forest, Decision Tree, and KNN, and Ensemble Classifiers such as AdaBoost and GradientBoost. The Phishing Testing Dataset is used for the experimental analysis. A total of 8955 cases and 32 attributes are considered. The Phishing Testing Dataset is available at []. Hard and Soft Voting classifiers are also used with high-accuracy classifiers. Individual and ensemble classifiers will be discussed in the subsections.

3.1 RANDOM FOREST:

The forest with many decision trees is created by the random forest algorithm. High detection accuracy is provided by large numbers of trees. The bootstrap method is used to create trees. In characteristics of the bootstrap algorithm and samples of the dataset are chosen at random and replaced to create a single tree. The random forest algorithm's randomly chosen features the best splitter for classification will be selected.

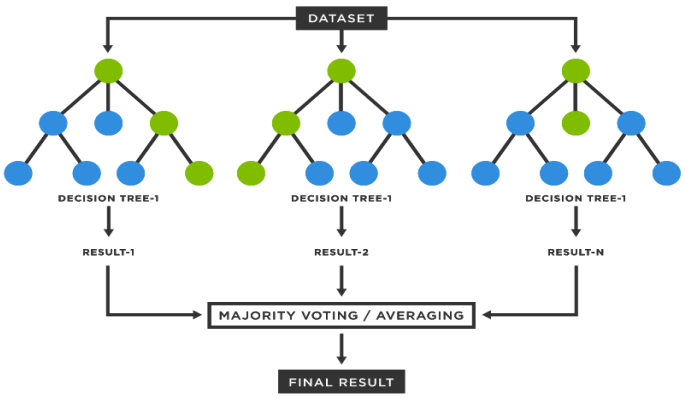

Fig -2: RandomForest

3.2 DECISION TREE:

A decision tree's job starts by calculating values such as the Gini index or information gain to find out attributes that are available for classification, which becomes the tree's root. This process keeps on repeating until a leaf node is found. In a tree representation, a decision tree provides a training model that predicts a target value or class. Each leaf node is a result label value and every middle node is some attribute from the dataset.

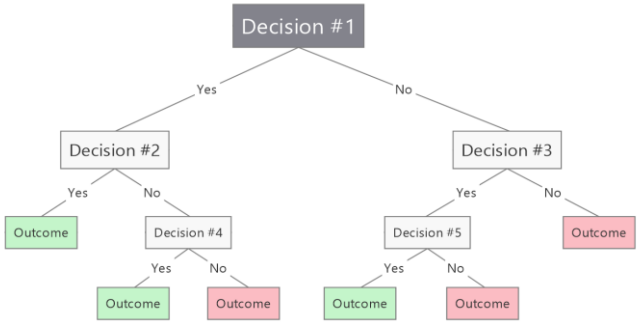

Fig -3: Decision Tree

3.3 KNN:

Among all the supervised learning techniques, the K-NN (K-Nearest Neighbor) is considered to be the simplest one. By assuming a resemblance between the new case/data and existing cases, the technique places the new data point in the group that best fits the available classifications. On the basis of similarity, in order to classify a new data point it stores all of the existing data. Thus employing the K-NN method, it is possible to promptly and reliably classify the new data into a relevant class. K-NN is more commonly utilized for classification problems, although there are cases when it can be employed for regression issues as well. Since the K-NN algorithm is a non-parametric algorithm, it does not rely on any underlying assumptions. The K-NN technique is also sometimes called a lazy learner method since it does not instantly learn from the training set; rather, it saves the dataset and applies a function to it when it has to categorize the data.

3.4 ENSEMBLE LEARNING:

A supervised machine learning technique called ensemble learning employs numerous learning algorithms to produce a single, ideal predictive model. To increase the model's capacity for prediction, it incorporates a variety of supervised learners. Techniques for ensemble learning that are frequently employed include bagging, boosting, and stacking. Ensemble classifiers categorize new samples using weights or majority voting and are built from a number of base (weak) classifiers. Different algorithms use base learners, often known as weak learners. As an illustration, consider the various base individual learners, such as K-NN, SVM, Bayesian, Decision Trees, etc. There are various tuning factors for base learners. Decision trees are the fundamental learning algorithm, the same as in the Random Forest ensemble. Weak learners are often referred to as base learners. Weak students frequently but irregularly record with little accuracy. When learning the connections between input and output, it is lacking. Weak

learners are combined in ensemble learning to achieve high accuracy. Based on the approach, the algorithm is chosen.

3.4.1 ADABOOST:

Adaptive boosting, often known as AdaBoost, is a commonly used boosting strategy that aims to turn multiple poor classifiers into one potent classifier. Only a single classifier is not often enough to predict the result with decent accuracy. By aggregating numerous weak classifiers and allowing each one to gradually learn from the incorrectly classified items of the others, a strong model can be created.

Considering a dataset with N points, or rows, in our dataset.

$$x_i \in \mathbb{R}^n, y_i \in \{-1, 1\}$$

In this case, n is the total number of classes in the dataset. The group of data points is x. y is our result variable which the data will be classified into. Each tuple is given a weight to prioritize some over others. When the assigned weights are high, those tuples or data points are taken into consideration for the next pass.

'w' which is the initial weighted sample will be the same for all the data points:

$$w = 1/N \in [0, 1]$$

Total error can be calculated as:

$$\alpha_t = \frac{1}{2} ln \frac{(1 - TotalError)}{TotalError}$$

Using this error weights can be updated,

$$w_i = w_{i-1} * e^{\pm\alpha}$$

The two possible outcomes for an alpha, positive or negative, show:

When some data is correctly classified in that case the value of alpha becomes positive. So we decrease its weight to reduce its priority.

When the result is incorrect, the value of alpha becomes negative. Here we increase the weight so this particular tuple has a high chance of being selected for training.

3.4.2 GRADIENTBOOST:

In gradient boosting, the error of the previous classification is improved in the next one.

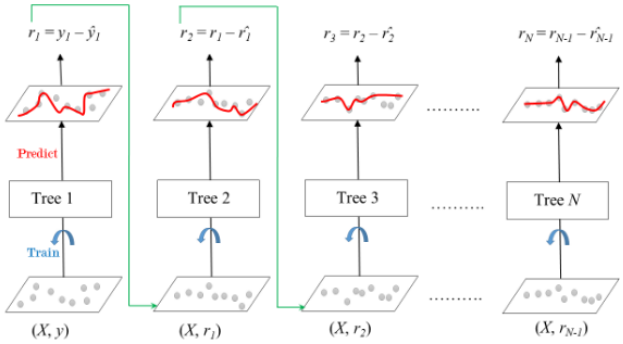

Fig -4: GradientBoost

The ensemble consists of N trees. The matrix X and labels Y are first used as training parameters for the first Tree. The residual errors r1 are determined using the labeled classifications y1. Then, Tree2 is trained with the r1 and X from Tree1 , and so on. Using the predicted outcomes r1, the residual r2 is determined. This goes on until every tree present is fully trained.

This makes use of a critical parameter known as shrinkage.

A forecast from a tree in an ensemble is said to have shrunk when it is multiplied by eta, which has values from 0 to 1. This is referred to as shrinking. To reach a certain level of model performance, the estimator-to-eta ratio must be balanced; a decrease in the learning rate necessitates an increase in the number of estimators. Now when every tree has been trained every prediction is possible.

$$y(pred) = y1 + (eta * r1) + (eta * r2) + \ldots\ldots + (eta * rN)$$

4. RESULTS AND DISCUSSION:

The efficiency of the ensemble classifier was assessed using a variety of evaluation metrics, including accuracy, recall, MCC, precision, and f-measure. These metrics are obtained from the Table 1 representation of the confusion matrix. The terms "TN" and "FN" stand for "amount of genuine websites detected as genuine," "TP" stands for "amount of phishing websites classified which are actually phishing websites," and "FP" stands for "amount of genuine sites wrongly detected as phishing websites." Table 2 shows the accuracies of different algorithms used.

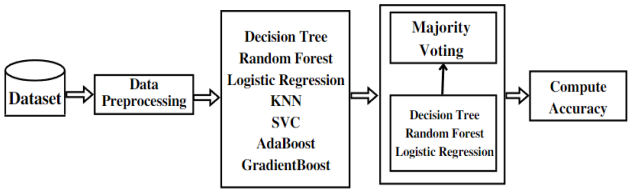

Fig -5: Process Flow Diagram of the Voting Classifier

Table-1: Confusion Matrix

	Predicted Phishing Websites	Predicted Legitimate Websites
Phishing Websites	**TP**	**FN**
Legitimate Websites	**FP**	**TN**

Table-2: Results obtained by Voting Classifier

Sr. No.	Metric	Value
1.	PPV or Precision	0.96
2.	Sensitivity or Recall	0.95
3.	Accuracy	0.969849
4.	MCC	0.9410
5.	F1 Score	0.97

Table-3: Accuracy obtained by various ML Techniques

Sr. No.	Metric	Value
1.	Random Forest	0.968638
2.	Logistic Regression	0.931656
3.	Decision Tree	0.951878
4.	K-Nearest Neighbor	0.649796
5.	Support Vector Machine	0.719113
6.	AdaBoost	0.943144
7.	Gradient Boost	0.953417
8.	Voting (Hard)	0.969849
9.	Voting (Soft)	0.968174

5. CONCLUSION:

Data from phishing websites can be classified using machine learning techniques, which are often employed. Although several methods for classifying the data on phishing websites have been created, there are still many difficulties, such as accuracy. We put up a strategy for categorizing data from phishing websites to address this. In order to categorize the data from phishing websites, an ensemble model is built in this work utilizing a voting classifier. Voting Classifier has decent classification accuracy. A variety of indicators are employed to gauge the model's effectiveness. In order to categorize the data from phishing websites, the suggested model is contrasted with other methods already in use. These two heterogeneous classifiers were combined, and the result was an exceptional accuracy of 96.98%. The findings demonstrate that the proposed strategies are superior to a single classifier from every angle. Users may successfully identify phishing websites with the suggested ensemble learning technique.

REFERENCES

[1] Atharva Deshpande, Omkar Pedamkar, Nachiket Chaudhary, Dr. Swapna Borde, 2021, Detection of Phishing Websites using Machine Learning, INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT) Volume 10, Issue 05 (May 2021),

[2] Dutta AK (2021) Detecting phishing websites using machine learning techniques. PLoS ONE 16(10): e0258361. https://doi.org/10.1371/journal.pone.0258361

[3] *Phishing website detection using machine learning algorithms - IJCA*. (n.d.). Retrieved October 26, 2022, from https://www.ijcaonline.org/archives/volume181/number23/mahajan-2018-ijca-918026.pdf

[4] M. Rastogi, A. Chhetri, D. K. Singh and G. Rajan V, "Survey on Detection and Prevention of Phishing Websites using Machine Learning," 2021 International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), 2021, pp. 78-82, doi: 10.1109/ICACITE51222.2021.9404714.

[5] S. Parekh, D. Parikh, S. Kotak, and P. S. Sankhe, "A New Method for Detection of Phishing Websites: URL Detection," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, vol. 0, no. Icicct, pp. 949–952.

[6] L. MacHado and J. Gadge, "Phishing Sites Detection Based on C4.5 Decision Tree Algorithm," in 2017 International Conference on Computing, Communication, Control and Automation, ICCUBEA 2017, 2018, pp. 1–5.