Phishing Website Detection Using Machine Learning Algorithms

Manjari Jayan
Department of Computer Applications
Amal Jyothi College Of Engineering
Kanjirappally, India
manjarijayan2024@mca.ajce.in

Jetty Benjamin
Department of Computer Applications
Amal Jyothi College Of Engineering
Kanjirappally, India
jettybenjamin@amaljyothi.ac.in

***Abstract*— The complexity of phishing assaults has increased, making it exceedingly challenging for the typical person to discern the legitimacy of an email message link or website. Nowadays, phishing efforts are a prevalent and successful tactic used by cybercriminals. In this section we provide an intelligent solution that reduces security risks for people and businesses by giving users an interface to determine whether a URL is phishing or authentic. The foundation of the system is machine learning, namely supervised learning. We chose the Gradient Boosting method because of its excellent classification performance.**

Our goal is to develop a better classifier by evaluating the characteristics of phishing websites and selecting the optimal combination of them to train the classifier with. As a result, our paper has 97.4% correctness and a total of 32 characteristics.

Keywords*— *Gradient Boosting, phishing, machine learning techniques

I. INTRODUCTION

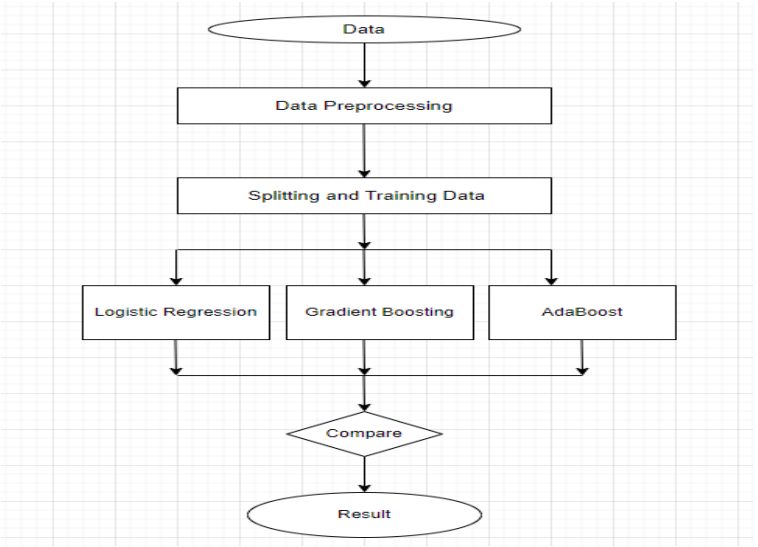
A phishing website is a well-known technique used by attackers to imitate trusted URLs and online platforms. Phishing is a process in which people construct fake websites that look like legitimate ones in order to steal sensitive or private information from unsuspecting customers. Attackers frequently utilize websites to deceive people into clicking on them in order to steal sensitive information such as login credentials, passwords, credit card numbers, and other personal and private information. These phishing attempts are also profitable for the perpetrators. They are mostly concerned with e-commerce applications, electronic payment systems, and online banking platforms. Phishing attacks are difficult to detect because they exploit the vulnerabilities of their targets. Therefore, it is critical to improve the mechanisms for detecting and preventing phishing attacks.

The "blacklist" strategy is one method for identifying phishing websites. This method entails including URLs that have been blacklisted by the antivirus database based on their Internet Protocol (IP) addresses. Attackers use devious methods to deceive people, such as obfuscating URLs and other simple tricks. For instance, they may employ flux, where proxies are automatically constructed to host the website, or they may use an algorithm to generate new URLs, among other approaches. The main drawback of this strategy is its inability to detect phishing assaults that begin immediately.

II. LITERATURE SURVEY

Phishing is the illegal act of impersonating a trustworthy source via an electronic message in order to steal private information such as usernames, passwords, and credit card details. [1]. To improve model precision, the suggested strategy employs a boosting technique that makes use of classifiers. [4]. To improve the accuracy of detecting phishing websites, we combine feature selection techniques with a learning strategy that makes use of regression and classification models. [2]. Machine learning can be used to detect phishing websites by evaluating their traits and attributes. [3]. Literature on the detection of phishing attacks [5].

III. METHODOLOGY



1. DATA SET DESCRIPTION

Features taken from URLs to categorize websites as phishing or non-phishing are included in the dataset that is provided. The features fall under the following categories:

a. Features related to the address bar

b. Features related to domains

c. Features that utilize JavaScript and HTML

a. Features related to the address bar

•URL Domain

It is necessary to extract and consider certain properties of the URL domain as features. The following choices have been made for this project out of all the available alternatives.

- Using IP

IP addresses included in URLs could be a sign of information theft attempts. If the domain portion of the URL contains an IP address, this is regarded as a sign of phishing activity and is given a value of -1. If an IP address is absent, the domain is deemed valid and is

awarded a rating of 1.

- The "@" sign in a URL

The system checks to see if the ", @" sign is present in the URL. When the "@" symbol occurs in a URL, the browser should usually disregard anything that comes before it; the real address usually appears after it. If the URL contains the "@" sign, the corresponding value for this feature is set to 1 (indicating phishing); otherwise it is assigned a value of -1 (indicating legitimacy).

- Short URL

Phishing scammers frequently use URL shortening services to make phishing URLs shorter. Their purpose is to direct consumers to websites.

- Prefix, suffix (-)

Determining whether the URL's domain contains the symbol ".". Dashes are often not used in legitimate URLs. Phishers frequently append prefixes or suffixes separated by dots to domain names to trick users into believing they are accessing a website.

- Redirection "//" in URL

It determines whether the URL contains the symbol "/". If the URL path contains any "//" characters, the user will be routed to a website.

- Favicon

The webpage about websites is pretty appealing. When it appears in the address bar, the webpage seems to be a phishing attempt.

- Https

This binary characteristic specifies whether the website uses HTTPS to communicate securely. A number of -1 indicates that the website does not employ HTTPS, whereas a value of 1 indicates that it does. HTTPS is frequently linked with genuine websites, however the absence of HTTPS might be an indication of phishing websites.

b. Features related to domains

Data extraction is yet another domain-specific function.

- DNS Record

When the DNS record for a phishing website is empty or cannot be detected, the rating for this feature is set to 1 (indicating phishing). Otherwise, it is assigned a value of -1 (indicating legitimacy).

- Web Traffic

This feature assesses a website's popularity by examining the number of visitors and the number of pages they view.

- Age of Domain

Phishing attempts typically only have a short lifespan. For this project, any domain must have been active for at least 12 months. In this context, the time between the domain's introduction and expiration is simply referred to as the domain's age.

c. Features that utilize JavaScript and HTML

The dataset also makes use of JavaScript and HTML-based features.

- IFrame Redirection

You can embed another website inside this one by utilizing the HTML element IFrame. The "iframe" tag enables phishers to hide or remove frame boundaries while engaging activities. Phishers instruct web browsers to differentiate themselves, in this case by utilizing the "frameborder" feature.

- Status Bar Custom

JavaScript can be used by scammers to fool viewers into believing that there is a URL in the status bar. To ascertain whether this feature is being utilized, one must scrutinize the website's source code, paying particular attention to the ", on Mouseover" event and its possible influence on altering the status bar.

- Disabling Right Click

Phishers frequently utilize JavaScript to deactivate features, which makes it difficult for users to read or save the source code of the website. The guidelines for utilizing "On Mouseover" to hide links are the same here. We can still check the webpage's source code to see if right clicking has been disabled by looking for the event "event.button==2".

- Website Forwarding

The frequency of website redirections can be used to distinguish between phishing and trustworthy websites. Our dataset indicates that trustworthy websites have never been redirected more than once.

2. DATA PREPROCESSING

Data processing includes tasks like feature extraction, normalization, cleansing, selection, transformation, and more. A training dataset is the final result of data preparation. The interpretation of the processing results can be affected by the preprocessing of the data. Data cleaning might involve activities like completing in any gaps in the dataset, eliminating information, finding and eliminating abnormalities or outliers, and fixing any discrepancies or incompatibilities. Data integration is one strategy that includes integrating databases or data sets. The process of collecting and standardizing data in order to measure a certain set of data is known as data transformation. The process of preparing data for analysis includes a number of processes, including eliminating any unnecessary information, picking particular cases or examples, and selecting the most pertinent attributes. Through the application of data reduction techniques, we are able to get study results from a substantially smaller condensed form of the dataset.

3. TRAIN-TEST SPLIT

To efficiently use the training dataset for detecting phishing websites in the testing dataset, our dataset—which consists of 11,054 instances and 31 features—is divided into two subsets: the testing set and the training set. Of these, 30 features are independent features and 1 dependent feature. In order to facilitate training and comprehension of the data, 30% of the data for the testing set is examined.

4. MACHINE LEARNING ALGORITHMS

Three different machine learning classification models are using here:

a. Gradient Boosting Classifier

Gradient boosting classifiers are a type of machine learning algorithms that build a strong predictive model by combining several weak learning models. Typically, decision trees are employed in gradient boosting. The use of boosting algorithms is essential when managing the bias-variance trade-off. In contrast to bagging algorithms, which solely manage high variance in a model, boosting manages both variance and bias and is said to be more efficient.

b. Decision Tree Classifier

The decision tree is among the most widely used machine literacy algorithms. It is a supervised learning-based classification model. It's easy to comprehend and implement. Its structure is tree-like, with an internal node representing the properties of the dataset, branches representing the decision-making process, and a leaf node indicating the outcome. The selection or test is guided by the characteristics of the provided dataset.

c. Logistic Regression

In order to visually depict data and highlight the link between a variable and one or more independent variables of nominal ordinal, interval, or ratio level nature, logistic regression is frequently used in prediction research. It is frequently used to predict the likelihood of a binary response by taking into account one or more variables. The usual application of logistic regression is to predict outcomes with two values, like 1 or -1.

IV.IMPLEMENATATION AND RESULT

A dataset from Kaggle was used to develop different machine learning classifiers for the purpose of detecting phishing websites. The model development and analysis were carried out using the Jupyter environment for Python. For analysis, the 11,054 instances and 32 features of the dataset were loaded into a Pandas dataframe. The dataset was examined in order to determine its structure and properties. A heatmap was used to assess feature correlation, which aided in detecting potential correlations between features, as shown in Fig. 1.

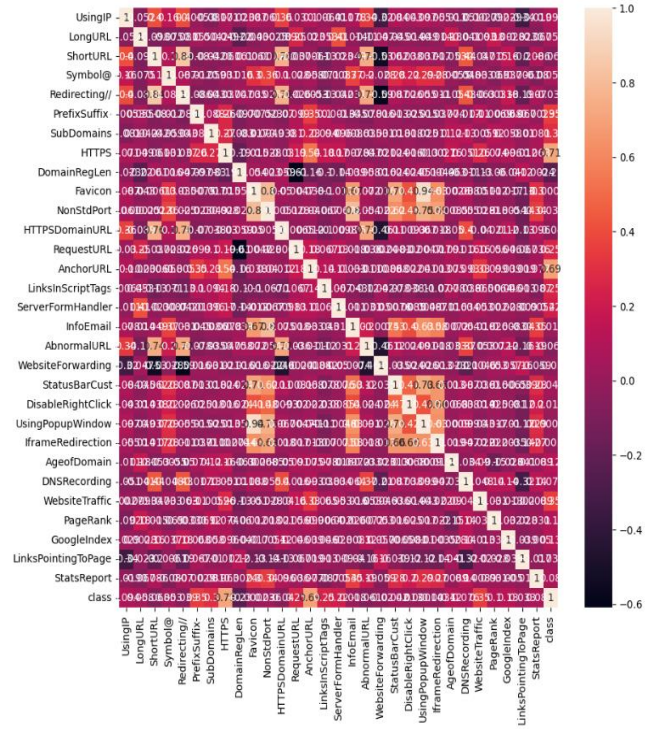


Fig. 1. Correlation heatmap

The distribution of phishing websites was represented by a pie chart, which showed the dataset's ratio of phishing to non-phishing cases, as shown in Fig. 2.

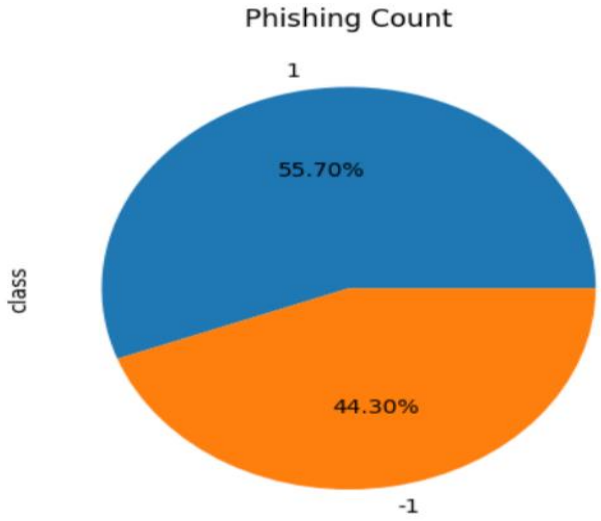


Fig. 2. Phishing Count in pie chart

Independent and dependent features were separated as part of the preprocessing of the dataset. An 80-20 ratio was used to divide the data into training and testing sets. Many metrics were calculated in order to assess the effectiveness of different classification models, such as Gradient Boosting Classifier, Decision Tree, and Logistic Regression. These metrics comprised precision, recall, F1-score, and accuracy. The model's hyperparameters were adjusted for peak performance.

The phishing website detection model is evaluated and trained using a variety of classifier and ensemble methodologies in order to guarantee the best degree of accuracy. Following the completion of each algorithm's results, each will indicate its estimated accuracy. Every algorithm is evaluated against other algorithms to determine which provides the highest accuracy rate, as shown in Table 1.

Classifiers	Accuracy	f1_score	Recall	Precision
Gradient Boosting Classifier	0.974	0.977	0.994	0.986
Decision Tree	0.960	0.964	0.991	0.993
Logistic Regression	0.934	0.941	0.943	0.927

Table 1. Comparison Table

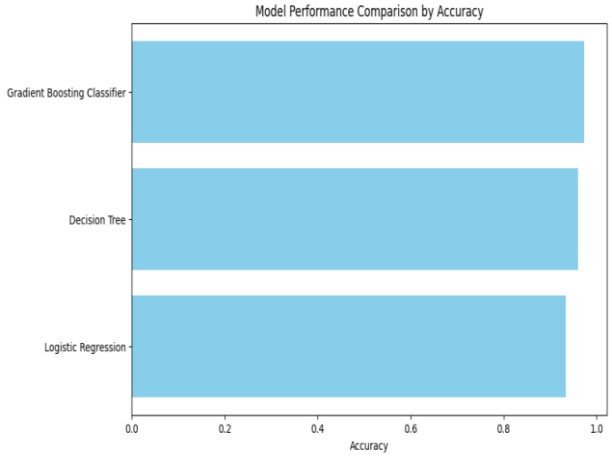


Fig. 3. Comparison of all ML algorithms

Among the tested models, the Gradient Boosting Classifier, with a particular configuration of max_depth and learning_rate, showed the highest accuracy. Furthermore, the visualization of feature importance enabled us to identify the most significant features in the model's predictions, as shown in Fig. 3.

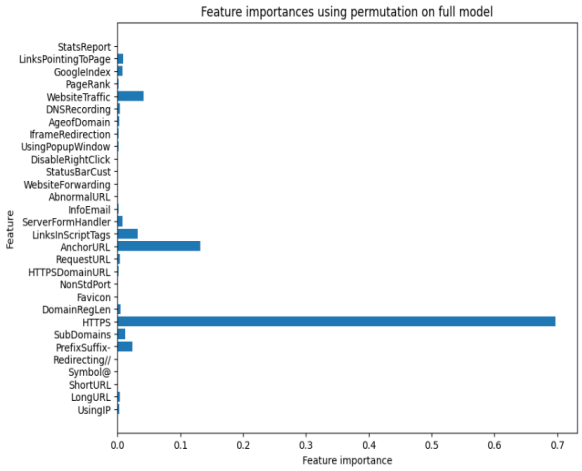


Fig. 3. Feature importance in the model

V. CONCLUSION

This work aims to enhance the detection of phishing websites by applying machine learning classification models. We achieve a detection accuracy of 97.4% using the Gradient Boosting Classifier whereas the Logistic Regression model yields an accuracy of 93.4%. Phishing on the internet carries some danger. The method created can be used as a resource for identifying websites and may stimulate the creation of more robust anti-phishing defenses in the future.

V. REFERENCES

1. Alyssa Anne Ubing, Syukrina Kamilia Binti Jasmi, Azween Abdullah, NZ Jhanjhi, Mahadevan Supramaniam “Phishing Website Detection: An Improved Accuracy through Feature Selection and Ensemble Learning”, 2019.
2. Mrs.Vaneeta M, Pratik N N,Prajwal D, Pradeep K S, Suhas Kakade K “Detection Of Phishing Websites Using Machine Learning Techniques”, 2020.
3. Ammar Odeh , Ismail Keshta and Eman Abdelfattah Adaptive Boosting Approach”, 2021.
4. Amani Alswailem, Bashayr Alabdullah, Norah Alrumayh, Aram Alsedrani, “Detecting Phishing Websites Using Machine Learning”,2019.
5. Lizhen Tang, Qusay H. Mahmoud, “ A Deep Learning-Based Framework for Phishing Website Detection”, December 2021