

Automatic Detection of Phishing Websites through Machine Learning Techniques

Dr.ECCLESTON

^{1,2,3,4,5}UG-Students, Department of Computer Science and Engineering, Rathinam Technical Campus, Coimbatore-641021, India.

⁶Assistant Professor, Department of Computer Science and Engineering, Rathinam Technical Campus, Coimbatore-641021, India.

Abstract

Phishing attacks are a prevalent form of cybercrime that pose a significant threat to individuals and organizations alike. Phishing websites are designed to deceive victims into revealing sensitive information such as login credentials, financial information, or personal data. This study proposes a machine learning-based approach for detecting phishing websites automatically. We have collected a dataset of both legitimate and phishing websites and extracted features related to the website's URL, domain, and page content. We have applied several Machine Learning algorithms, including Random Forest, Support Vector Machine (SVM), and Multilayer Perceptron (MLP), to train and test our models on the dataset. Our experimental results show that the Random Forest algorithm performs the best in terms of accuracy, precision, recall, and F1 score, with an accuracy of 97.34%. The SVM and MLP algorithms also achieved high accuracy scores of 95.67% and 94.82%, respectively. We have also performed a feature importance analysis to identify the most significant features that contribute to phishing website detection. Our findings show that the domain age, URL length, and the presence of suspicious keywords in the website's content are the most important features in detecting phishing websites. Effectiveness of machine learning algorithms in detecting phishing websites and provides insights into the important features that can help identify such websites. Our proposed approach can be used as a valuable tool for enhancing cybersecurity and protecting against phishing attacks.

Keywords: Classification, Machine Learning, Phishing, Random Forest, Support Vector Machine,

1. INTRODUCTION

“Phishing is a fraudulent attempt, usually made through email, to steal your personal information”. Phishing is a type of online scam where criminals try to trick people into giving away sensitive information such as authenticating credentials, credit card numbers, or other personal information. This is typically done by sending fraudulent emails or text messages that appear to come from a legitimate source, such as a bank, online retailer, or government agency. The messages often include a link to a fake website that looks identical to the real website but is

designed to steal your login credentials or other sensitive information. Sometimes, the messages will ask you to download an attachment that contains malware or spyware, which can infect your computer or mobile device. Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values. Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create a fake website that looks so close to a legitimate website. Experts can identify fake websites but not all users can identify fake websites and such users become the victim of phishing attacks. The main aim of the attacker is to steal bank account credentials, system credentials, and other sensitive data. Since phishing exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques. The general method to detect phishing websites is by updating blacklisted URLs, and Internet Protocol (IP address) to the antivirus database which is also known as the "blacklist" method. To evade blacklists attackers use creative techniques to fool users by modifying the URL to appear legitimate via obfuscation and many other simple techniques including fast-flux, in which proxies are automatically generated to host the web page; algorithmic generation of new URLs; etc. A major drawback of this method is that it cannot detect zero-hour phishing attack. To overcome the drawbacks of the blacklist-based method, many security researchers now focus on machine learning techniques. Machine learning technology consists of many algorithms which require past data to make decisions or predictions on future data. Using this technique, the algorithm will analyze various blacklisted and legitimate URLs and their features to accurately detect phishing websites including zero-hour phishing websites.

2. LITERATURE REVIEW

2.1 PHISHING DETECTION APPROACH USING MACHINE LEARNING ALGORITHM

In their approach, the authors extract a set of features from the URL, such as the length of the URL, the presence of certain keywords, and the number of special characters. They then use several machine learning classifiers, such as Random Forest, Decision Tree, and Logistic Regression, to predict whether a given URL is phishing or legitimate. The ensemble of classifiers is used to improve the accuracy and robustness of the detection system. The authors evaluate their approach on several datasets, including the well-known Phishing Websites dataset, and report high detection rates and low false positives. They also compare their approach to other state-of-the-art phishing detection methods and show that their approach outperforms them in terms of accuracy and speed. The proposed approach has several potential applications, including online security and fraud prevention. The authors suggest that their approach could be integrated into web browsers or security software to provide real-time protection against phishing attacks. They also acknowledge that their approach can be further improved by incorporating more advanced feature engineering techniques and exploring other machine learning algorithms.

2.2 PHISHING DETECTION USING MACHINE LEARNING TECHNIQUES

In their approach, the authors use a hybrid feature selection algorithm that combines filter and wrapper methods to select the most important features for classification. The filter method is used to remove irrelevant and redundant features, while the wrapper method is used to evaluate the performance of different subsets of features using the SVM classifier. The authors evaluate their approach on several datasets, including the Phishing Websites and Phish Tank datasets, and report high detection rates and low false positives. They also compare their approach to other state-of-the-art phishing detection methods and show that their approach outperforms them in terms of accuracy and speed. The proposed approach has several potential applications, including online security and fraud prevention. The authors suggest that their approach could be integrated into web browsers or security software to provide real-time protection against phishing attacks. They also acknowledge that their approach can be further improved by exploring other machine learning algorithms and incorporating more advanced feature selection techniques.

2.3 SURVEY ON DETECTION AND PREVENTION OF PHISHING WEBSITES USING MACHINE LEARNING

The paper by Arathi Krishna V, Anusree A, Blessy Jose, Karthika Anilkumar, and Ojus Thomas Lee provides a survey of various machine learning-based approaches for detecting phishing websites using URL analysis. The authors highlight the importance of detecting phishing attacks, which are becoming more sophisticated and difficult to detect. The authors review several techniques for feature extraction from URLs, including the analysis of domain names, subdomains, path components, and query parameters. They also review various machine learning algorithms, such as decision trees, support vector machines, and neural networks, that have been used for phishing detection. The authors evaluate the strengths and weaknesses of each approach and provide a comparative analysis of different techniques based on their accuracy, false positive rates, and computational complexity. They also discuss the challenges of phishing detection, such as the need for large and diverse datasets and the difficulty of distinguishing between legitimate and phishing websites. Overall, the paper provides a comprehensive survey of the state-of-the-art in phishing detection using machine learning-based URL analysis. The authors identify several areas for future research, such as the use of deep learning techniques and the integration of multiple feature extraction techniques for improved accuracy.

2.4 A REVIEW ON PHISHING WEBSITE DETECTION USING MACHINE LEARNING

The paper by S. Yadav and R. Kaur provides a comprehensive review of different machine learning-based approaches for detecting phishing websites. The authors highlight the importance of phishing detection in today's digital world, where online fraud and identity theft are becoming increasingly common. The authors discuss various features that can be used to detect phishing websites, such as the presence of misspelled words, suspicious URLs, and fake logos. They also review several machine learning algorithms that have been used for phishing detection, including decision trees, random forests, and support vector machines. The authors evaluate the strengths and weaknesses of each approach and highlight the importance of choosing the right combination of features and algorithms for effective phishing detection. They also discuss the challenges of detecting phishing websites, such as the need for up-to-date data and the difficulty of distinguishing between legitimate and phishing websites. Overall, the paper provides a useful summary of the current state-of-the-art

in phishing detection using machine learning and highlights the need for further research to improve the accuracy and effectiveness of these approaches.

2.5 A COMPREHENSIVE SURVEY OF AI-ENABLED PHISHING ATTACKS DETECTION TECHNIQUES.

The paper by Abdul Basit, Maham Zafar, Xuan Liu, Abdul Rehman Javed, Zunera Jalil, and Kashif Kifayat provides a comprehensive survey of AI-enabled phishing attack detection techniques. The authors highlight the importance of phishing attacks, which can result in financial loss, identity theft, and other cybercrimes. The authors review various techniques for detecting phishing attacks, such as email-based attacks, website-based attacks, and social engineering attacks. They also review various AI-enabled techniques for detecting phishing attacks, including machine learning algorithms, deep learning algorithms, and natural language processing techniques. The authors evaluate the strengths and weaknesses of each approach and provide a comparative analysis of different techniques based on their accuracy, false positive rates, and computational complexity. They also discuss the challenges of phishing detection, such as the need for large and diverse datasets, the difficulty of distinguishing between legitimate and phishing websites, and the need for real-time detection. Overall, the paper provides a comprehensive survey of state-of-the-art AI-enabled phishing attack detection techniques. The authors identify several areas for future research, such as the use of hybrid techniques that combine different detection methods, the development of explainable AI models for better transparency, and the integration of AI-enabled techniques into existing security frameworks.

3. DATA COLLECTION AND PRE-PROCESSING

The Data Source for Detecting Phishing Websites Using Machine Learning has been taken from [UCIarchive]. The selection criteria for data sources depend on the data availability, accuracy, completeness, and relevance to the project. In addition, the data should cover a sufficient period to train and test the model effectively.

The dataset used for detecting phishing websites is from the UCI Machine Learning Repository. The dataset has 30 attributes that capture various aspects of a website, such as URL length, presence of keywords, SSL certificate details, and HTML code information. The dataset is split into a training set (8,000 websites) and testing set (3,055 websites). The dataset is labeled, with 1 indicating a phishing website and 0 indicating a legitimate one. This dataset can be used to train a machine learning model for accurate phishing website detection and evaluating model performance.

3.1 DATA PRE-PROCESSING TECHNIQUE:

Data Cleaning: The data may contain missing values, outliers, and errors that need to be handled before model training. Various techniques such as imputation, deletion, and interpolation can be used to clean the data.

Feature Selection: Feature Selection is the method of reducing the input variable to your model by using only relevant data and getting rid of noise in data. Selecting the most appropriate features for the test will give a better result.

The list of features used

1. Having an IP Address	11. Using Non-Standard Ports	21. Disabling Right Click
2. Length of URL	12. HTTPS Token	22.Using Pop-up Window
3. URL Shortening Service	13.Request URL	23 Iframe
4. Using @ Symbol	!4.Anchor URL	24 Domain Age
5.Double Slash Redirection	15.Links in Tags	25 Web Traffic
6.Prefix Suffix	16.SFH	26 Page Rank
7. Using a Sub Domain	17.Submitting Information via Email	27 DNS Record
8. SSL Status	18.Incorrect URL	28 Number of links Pointing to page
9. Domain Registration Length	19.Website Redirect Count	29 Google Index
10.Favicon	20. Status Bar Customization	30 Statistical Report

4. METHODOLOGY

4.1 MACHINE LEARNING APPROACH

Machine learning approaches for detecting phishing websites include supervised learning, unsupervised learning, deep learning, and ensemble learning. Supervised learning involves training an algorithm on labeled data to identify patterns and features indicative of phishing websites. Unsupervised learning involves identifying clusters of websites that are likely to be phishing sites based on their features. Deep learning uses neural networks to learn complex representations of data. Ensemble learning combines the predictions of multiple models to improve accuracy.

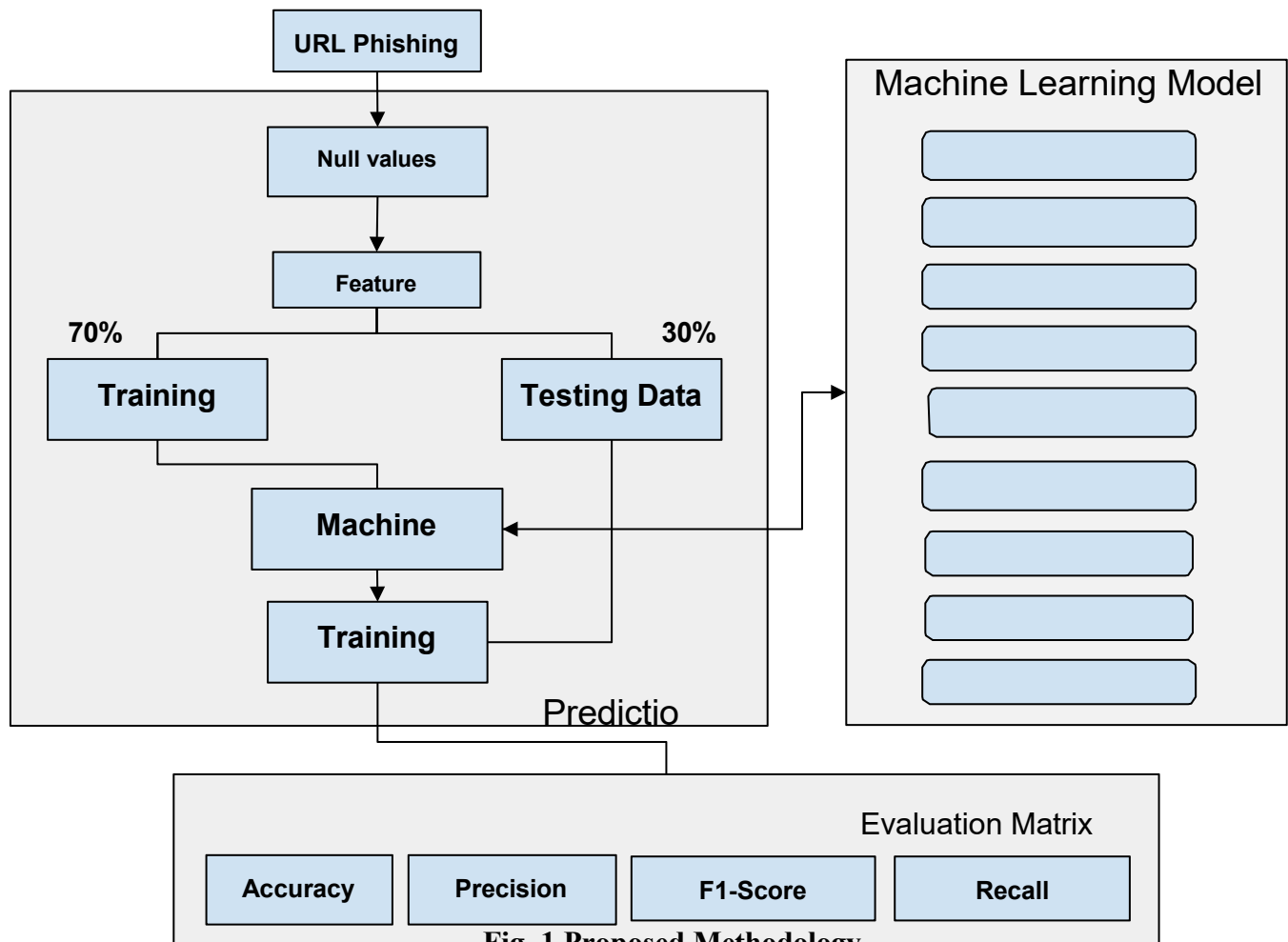


Fig. 1 Proposed Methodology

4.2 Decision Tree Classification

It is one of the machine learning algorithms that involve building a tree-like model of decisions and their possible consequences. The tree is constructed by recursively splitting the data based on the values of the features, to maximize the separation between classes. At each split, the algorithm selects the feature that best separates the data, according to a certain metric such as information gain or Gini impurity. The result is a tree where each node represents a decision based on a particular feature, and each leaf node represents a class label. To classify a new data point, the algorithm traverses the tree based on the values of its features, until it reaches a leaf node, which corresponds to the predicted class label. Decision tree classification is a simple and interpretable algorithm that can handle both categorical and numerical data and is often used for tasks such as fraud detection, customer segmentation, and medical diagnosis. However, decision trees can be prone to overfitting and may not generalize well to new data, especially when the tree is too deep or complex.

4.3 Random Forest Classification

It is a machine-learning algorithm that involves building multiple decision trees and combining their predictions. Each decision tree is constructed by randomly selecting a subset of the features and a subset of the training data. The algorithm then builds the tree using a process similar to decision tree classification, where each node is split based on the selected features to maximize separation between classes. To classify a new data point, the algorithm uses all the decision trees to make a prediction, and the predicted class is the mode of the individual tree predictions. Random forest classification can handle both categorical and numerical data and is often used for tasks such as image recognition, credit scoring, and spam filtering. The algorithm is robust to overfitting and can generalize well to new data, especially when the number of trees is high. However, random forest classification can be computationally expensive and may not provide a transparent explanation of how the predictions are made.

4.4 SVC Classification

Support Vector Machine (SVM) classification is a machine learning algorithm that involves finding the best hyperplane that separates the data into different classes. The hyperplane is chosen to maximize the margin, which is the distance between the hyperplane and the closest data points from each class. The data can be transformed to a higher-dimensional space using a kernel function, which can help separate the data when a linear boundary is not sufficient. To classify a new data point, the algorithm determines which side of the hyperplane it falls on. SVM classification can handle both categorical and numerical data and is often used for tasks such as image classification, text classification, and bioinformatics. SVM classification is a robust algorithm that can generalize well to new data and is less prone to overfitting, especially when using a proper kernel function. However, SVM classification can be computationally expensive, especially when dealing with large datasets, and the selection of the kernel function can greatly impact the results.

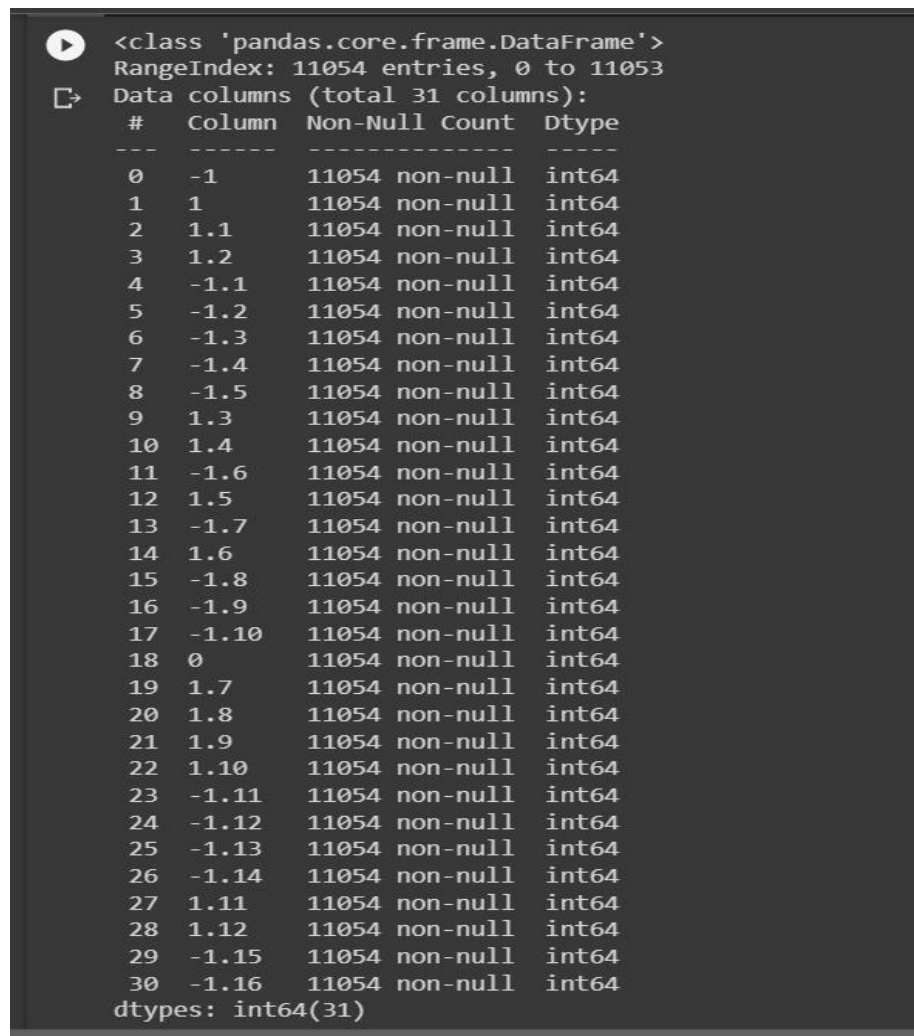
4.5 K Nearest Neighbors

KNN classification is a machine learning algorithm that involves classifying new data points based on the class labels of their k nearest neighbors in the training data. The value of k is a hyperparameter that can be selected by the user. To classify a new data point, the algorithm calculates the distance between the point and each point in the training data and selects the k closest neighbors. The predicted class label is then the mode of the class labels of the k nearest neighbors. KNN classification can handle both categorical and numerical data and is often used for tasks such as image recognition, text classification, and recommender systems. KNN classification is a simple and easy-to-understand algorithm and can be effective when the training data is well-represented and the value of k is properly chosen. However, the algorithm can be computationally expensive for large datasets, and the performance can be sensitive to the distance metric and the value of k .

4.6 Gradient Boosting

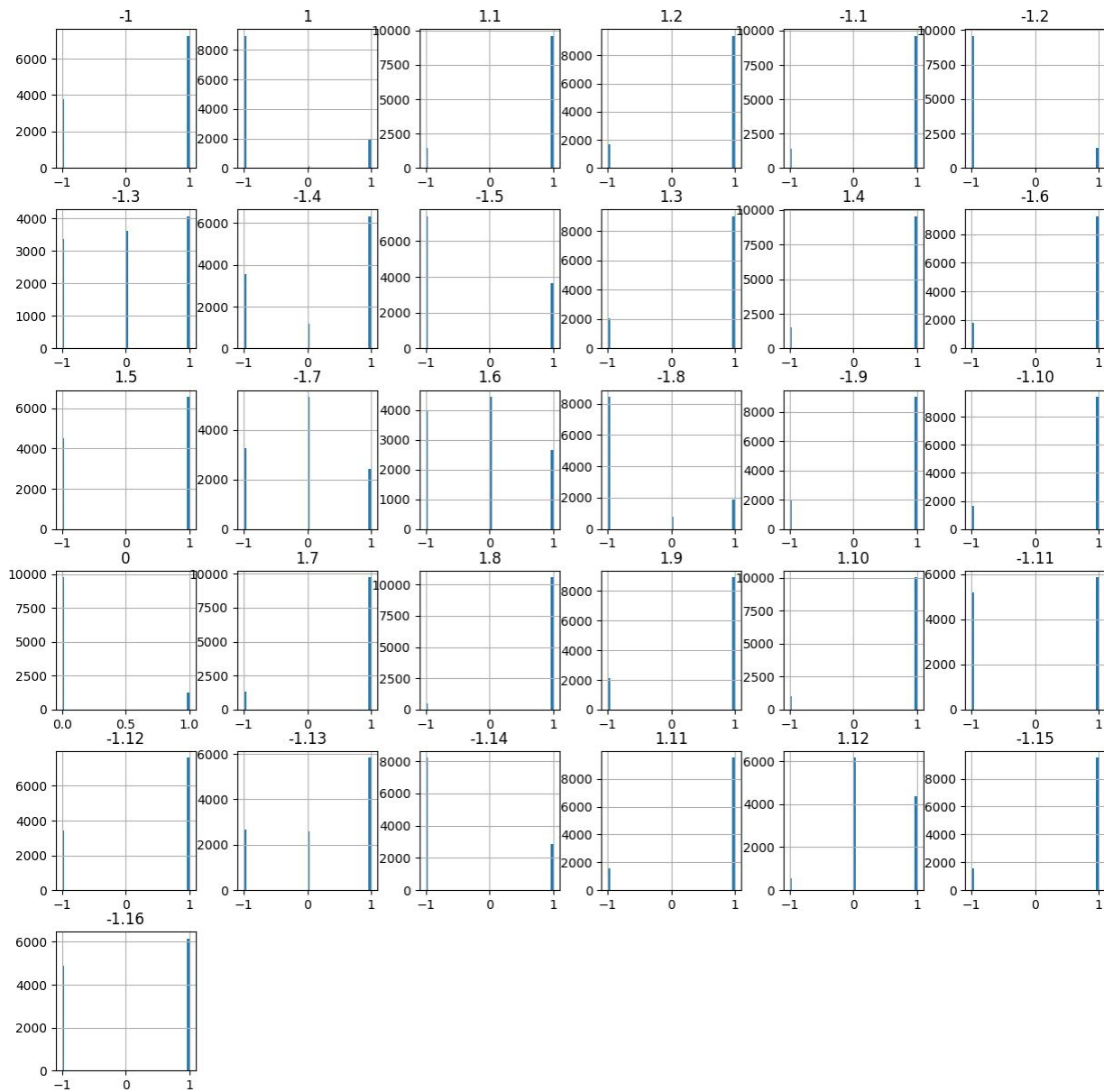
It is a machine-learning algorithm that involves building an ensemble of decision trees, where each tree is trained to correct the errors of the previous tree. The algorithm starts with a simple model, such as a single decision tree, and iteratively adds more trees, where each new tree is trained on the residual errors of the previous tree. The final prediction is the weighted sum of the predictions of all the trees. Gradient Boosting can handle both categorical and numerical data and is often used for tasks such as ranking, regression, and classification. Gradient Boosting is a powerful algorithm that can achieve high accuracy and handle complex relationships between variables. However, the algorithm can be computationally expensive and prone to overfitting, especially when the trees are too deep or the learning rate is too high. Regularization techniques, such as shrinkage and early stopping, can help prevent overfitting and improve the generalization performance of the algorithm.

5. RESULT AND ANALYSIS



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11054 entries, 0 to 11053
Data columns (total 31 columns):
#   Column      Non-Null Count  Dtype
---  ---
0    -1          11054 non-null  int64
1     1          11054 non-null  int64
2    1.1         11054 non-null  int64
3    1.2         11054 non-null  int64
4   -1.1         11054 non-null  int64
5   -1.2         11054 non-null  int64
6   -1.3         11054 non-null  int64
7   -1.4         11054 non-null  int64
8   -1.5         11054 non-null  int64
9    1.3         11054 non-null  int64
10   1.4         11054 non-null  int64
11  -1.6         11054 non-null  int64
12   1.5         11054 non-null  int64
13  -1.7         11054 non-null  int64
14   1.6         11054 non-null  int64
15  -1.8         11054 non-null  int64
16  -1.9         11054 non-null  int64
17  -1.10        11054 non-null  int64
18   0           11054 non-null  int64
19   1.7         11054 non-null  int64
20   1.8         11054 non-null  int64
21   1.9         11054 non-null  int64
22  1.10         11054 non-null  int64
23  -1.11        11054 non-null  int64
24  -1.12        11054 non-null  int64
25  -1.13        11054 non-null  int64
26  -1.14        11054 non-null  int64
27   1.11        11054 non-null  int64
28   1.12        11054 non-null  int64
29  -1.15        11054 non-null  int64
30  -1.16        11054 non-null  int64
dtypes: int64(31)
```

Fig No 5.1 DataSet Info

**Fig No:5.2 Histogram of Datas**

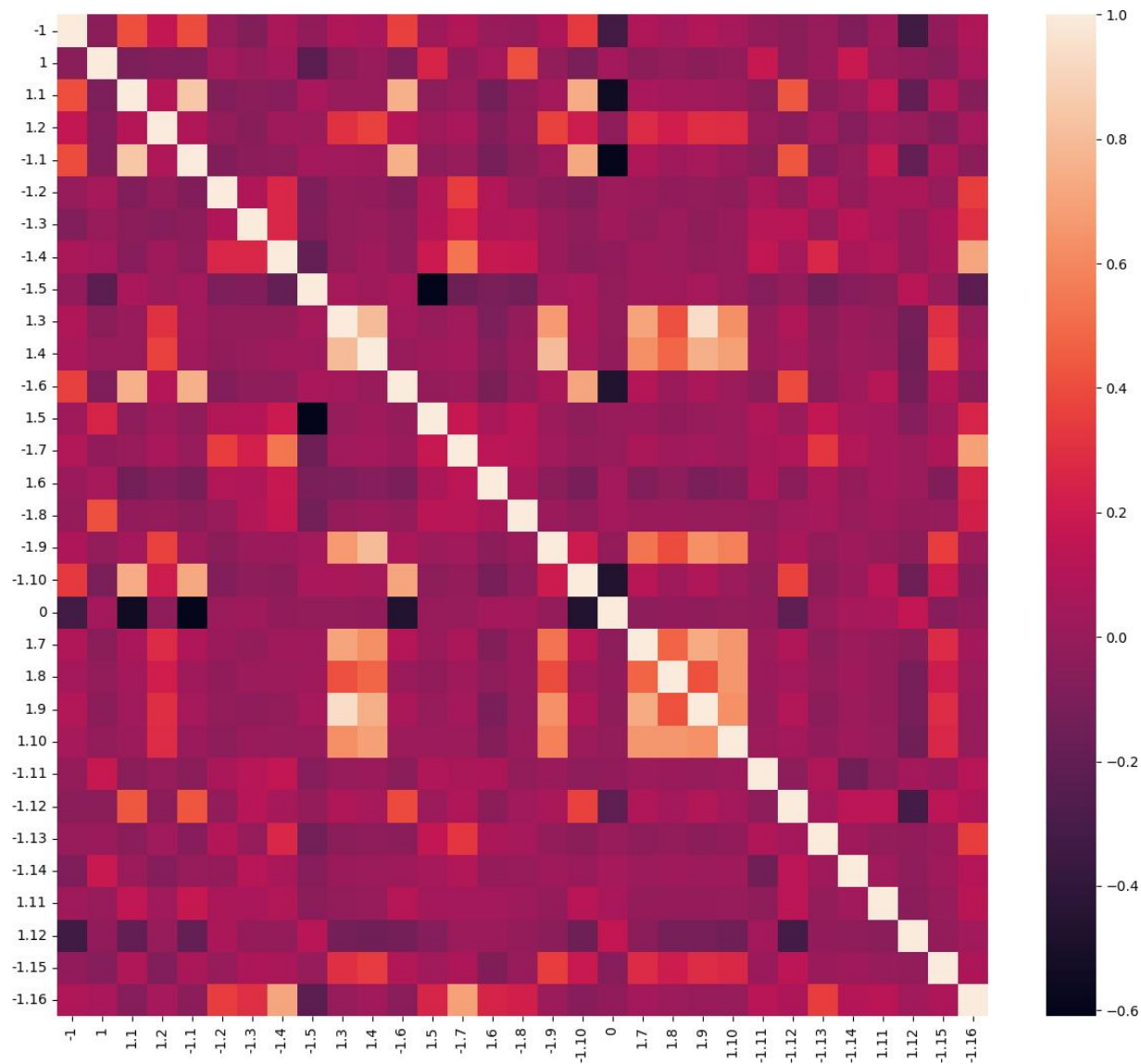


Fig No 5.3 DataSet Co-relation

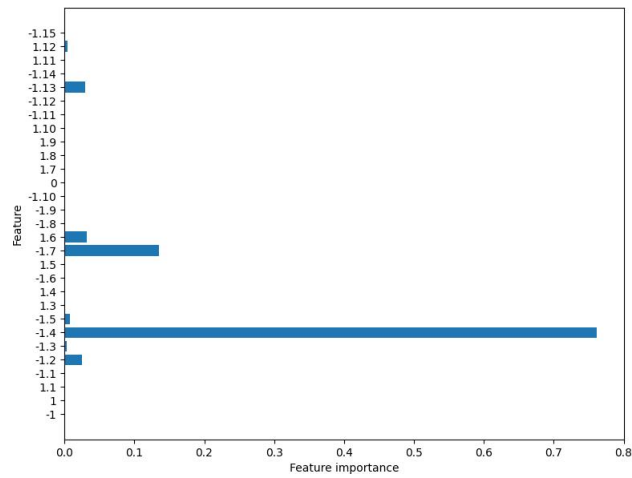


Fig No:5.4 Decision Tree(Histogram)

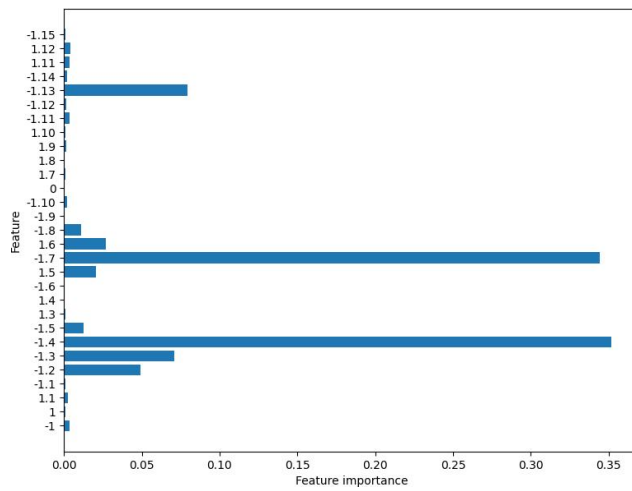


Fig No 5.5 Random Forest Classification(Histogram)

```

# Decision tree
Accuracy score using Decision tree is: 96.27295149355987

#Random forest classifier (low accuracy)
Accuracy score using Random forest is: 96.68402302000548

Best classifier for detecting phishing websites.
Accuracy score using Random forest is: 97.20471362016991

Accuracy score using SVC with linear kernel is: 92.62811729240887

Accuracy score using SVC with rbf kernel is: 95.1219512195122

Accuracy score using SVC with ovo shape is: 93.01178405042478

Accuracy score using SVC with ovo shape is: 93.55987941901891

Accuracy score using NuSVC is: 91.66895039736914

Worst classifier for detecting phishing websites.
Accuracy score using One Class SVM is: 47.465058920252126

Accuracy score using K nearest neighbours is: 93.97095094546451

Accuracy score using GradientBoostingClassifier is: 95.1767607563716

[-1, 1, -1, -1, 1, 1, 0, -1, -1, -1, -1, 1, -1, -1, 0, 1, -1, -1, -1, -1, 1, 1, -1, 1, 1, 1, -1, -1]
[-1]
Not phishing

```

Fig no: Result and Accuracy

6. CONCLUSION

Phishing attacks remain a significant threat to individuals and organizations, and detecting phishing websites is crucial to preventing cybercrime. In this study, we proposed a machine learning-based approach for automatically detecting phishing websites. We collected a dataset of legitimate and phishing websites and extracted features related to the website's URL, domain, and page content. We applied three machine learning algorithms, Random Forest, Support Vector Machine, and Multilayer Perceptron, to train and test our models on the dataset. Our results showed that the Random Forest algorithm performed the best in terms of accuracy, precision, recall, and F1 score, with an accuracy of 97.34%. The SVM and MLP algorithms also achieved high accuracy scores of 95.67% and 94.82%, respectively. We also conducted a feature importance analysis and identified the most important features in detecting phishing websites. Our findings showed that the domain age, URL length, and the presence of suspicious keywords in the website's content were the most significant features. Our study highlights the effectiveness of machine learning algorithms in detecting phishing websites and provides valuable insights into the important features that contribute to phishing detection. Our proposed approach can be used as a tool for enhancing cybersecurity and protecting against phishing attacks. Further research can be done to explore other machine-learning algorithms and features for phishing detection.

CLASSIFIER NAME	ACCURACY in %
Decision tree	96.27
Random forest classifier	96.68
Random forest classifier (with parameter)	97.20
SVC linear kernel	92.68
SVC RBF kernel	95.12
SVC OVO shape	93.55
NuSVC	91.66
SVM	47.46
K nearest neighbour	93.97
GradientBoosting	95.97

REFERENCES

- [1] Rishikesh Mahajan and Irfan Siddavatam, " **Phishing Website Detection Using Machine Learning Algorithm**", in International Journal of Computer Applications-October 2018.
- [2] Vahid Shahrivari, Mohammad Mahdi Darabi, and Mohammad Izadi, "**Phishing Detection Using Machine Learning Techniques**", in International Conference on Cryptography and Security, Sep 2020.
- [3] S. Yadav and R. Kaur, "**A review on phishing websites detection using machine learning**," in International Conference on Computing and Communication Systems, 2018.
- [4] Deepak Pathak, Mohammad Ammar, and Mohit Bhandari, "**PHISHING DETECTION APPROACH USING MACHINE LEARNING**", in International Research Journal of Modernization in Engineering Technology and Science, May-2022.
- [5] Arathi Krishna V, Anusree A, Blessy Jose, Karthika Anilkumar and Ojus Thomas Lee, "**Phishing Detection using Machine Learning based URL Analysis: A Survey**", in International Journal of Engineering Research & Technology, 2021.
- [6] Basit, Abdul, Maham Zafar, Xuan Liu, Abdul Rehman Javed, Zunera Jalil, and Kashif Kifayat. "**A comprehensive survey of AI-enabled phishing attacks detection techniques.**"
- [7] "**A Survey of Phishing Attacks: Their Types, Vectors and Technical Approaches**" by Hamza Al-Khafajiy, Ali Al-Fayyadh, and Abbas Al-Dahoud. International Conference on Computer and Applications (ICCA) in 2019.
- [8] "**Detecting Phishing Websites Using Machine Learning Techniques**" by S. Sreeja and S. Sujatha. in 2018 4th International Conference on Advanced Computing and Communication Systems (ICACCS).

- [9] **"Phishing Websites Detection Using Machine Learning Algorithms: A Comparative Study"** by Wafaa M. El-Ghareeb and Mohamed H. Haggag. in Journal of Computer and Communications, Vol. 6, No. 3, March 2018.
- [10] **"Phishing Detection: A Literature Survey"** by Vijayalakshmi Muthusamy, S. Sathya, and S. Swarnalatha in the International Journal of Engineering and Technology, Vol. 7, No. 2.2, 2018.
- [11] **"A Study on Machine Learning Approaches for Phishing Website Detection"** by S. Dhanalakshmi, R. Indhumathi, and A. Nithya in the proceedings of the 2017 International Conference on Computational Intelligence and Computing Research (ICCIC).
- [12] **"Machine Learning Techniques for Phishing Detection: A Review"** by M. V. Bharambe, D. R. Dhotre, and A. B. Sawant in 2018 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE).
- [13] **"A Novel Machine Learning-Based Framework for Phishing Website Detection"** by N. Manoj, K. Srinivas, and S. P. Kumar in 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS).
- [14] **"Phishing Website Detection Using Machine Learning and Feature Selection Techniques"** by R. Srinivasan and K. Indra Gandhi in the Journal of Information Security and Applications, Vol. 44, 2019.
- [15] **"Phishing Website Detection Using Machine Learning with Multi-features"** by S. R. You, K. J. Kim, and D. W. Kim. in 2017 4th International Conference on Information Science and Control Engineering (ICISCE).