PhishGuardian: Random Forest-based Phishing Website Detection and Reporting System

Rajesh N V
Dept. CSE
BMS Institute of Technology and Management,
Bengaluru
Karnataka – 560064
rajesh@bmsit.in

Ravi Kiran Poudel
Dept. CSE
BMS Institute of Technology and Management,
Bengaluru,
Karnataka - 560064
1by20cs147@bmsit.in

Ullas B K
Dept. CSE
BMS Institute of Technology and Management,
Bengaluru,
Karnataka - 560064
1by20cs209@bmsit.in

Ramchandra Paudel
Dept. CSE
BMS Institute of Technology and Management,
Bengaluru,
Karnataka - 560064
1by20cs145@bmsit.in

Amarnath Mahato
Dept. CSE
BMS Institute of Technology and Management,
Bengaluru,
Karnataka - 560064
1by20cs019@bmsit.in

Abstract: ***As social life and the Internet become more deeply integrated, people's perspectives on learning and work are changing, and this exposes us to an increasing number of serious security threats. The ability to recognize various network threats, particularly ones that haven't been seen before, is a critical problem that needs to be dealt with immediately. Phishing site URLs are designed to collect private information such as user identities, passwords, and online money transactions. Phishers use websites that are visually and semantically similar to authentic websites. Since many clients go online to access government services as well as financial services, there has been a significant increase in phishing threats and attacks over the past few years. As technology advances, phishing methods have accelerated, and this should be avoided by employing anti-phishing techniques to detect phishing. Machine learning is an authoritative tool that can be employed to combat phishing attacks. There are various methods or techniques used to identify phishing websites. Machine learning approaches for detecting phishing websites were previously proposed and implemented. The primary goal of our project is to extract relevant features from website URLs and implement the system efficiently and accurately. Our project employs the Random Forest algorithm, which has been further customized to provide more accurate predictions by varying the algorithm's input parameters and thus selecting the optimal default configuration to create the model, resulting in a 97% accuracy score.***

Keywords: ***Phishing, Random Forest (RF), URL, Prediction, Machine Learning (ML), Legitimate, Decision Tree (DT)***

I. INTRODUCTION

In the contemporary digital landscape, the proliferation of cybersecurity threats has become an issue that poses significant risks to both individuals and institutions. Of the various forms of cyberattacks, phishing stands out as being among the most ubiquitous, wherein attackers employ deceptive tactics to procure sensitive user information like passwords, credit card numbers, and usernames information by posing as legitimate entities. Phishing attacks can be launched through various channels, such as e-mail, messaging apps, instant messaging, social media, and even counterfeit websites that closely mimic authentic ones. Detecting such spurious websites can prove to be an arduous task, as attackers utilize a variety of different techniques to obfuscate their malicious intent and pass off their websites as genuine. Traditional signature-based methods are ineffective in recognizing new and previously unseen attacks, while manual examination of websites is both time-consuming and infeasible. Therefore, it is of the highest priority to have automated approaches to classify and identify fraudulent websites.

In our project, we have focused primarily on creating features that help to identify malicious URLs. ML algorithms are used on a categorical dataset comprising malware, phishing, defacement, and benign URLs. The inclusion of Internet Protocol (IP) addresses and suspicious words in URLs is highly effective in detecting fraudulent URLs. Overall, our project highlights the significance of feature engineering in creating robust and efficient utilization of the Random Forest algorithm for detecting malicious URLs.
By identifying the most effective lexical features, we can further enhance the accuracy and efficiency of these techniques and make it easier to detect and prevent cyber-attacks.

II. PROBLEM DESCRIPTION

The issue arises following a comprehensive examination of the machine learning-based classification method for identifying phishing websites. Our task is to design a system that minimizes time investment while enabling efficient differentiation between legitimate and phishing sites, thereby enhancing accuracy.

III. MOTIVATION

With the increasing growth in the technology field, there is an alarming growth of phishing attacks, which threaten digital identities, privacy, and financial security. Phishing, a prevalent cyber threat, targets individuals and organizations, compromising sensitive data and causing financial losses. Following the APWG, over 241,324 unique phishing attacks were reported throughout the first half of 2023 [1]. Anti-phishing technology safeguards digital identities, preserving trust in online interactions, and preventing identity theft. It also has an important part to play in preserving privacy and data security, mitigating financial losses, and providing insights to combat cybercriminals. ML-based approaches used for anti phishing technology have produced favorable outcomes when it comes to detecting phishing websites by analyzing their features and characteristics. This motivated us to build an RF-based system for the detection of phishing websites that improves the precision and effectiveness of the malware detection process and enhances the online security of individuals and organizations.

IV. LITERATURE SURVEY

We've reviewed some previous research works on fake URL detection using ML algorithms in this section. The system described classifies the web addresses (URLs) as either phishing or non-phishing or a fake website or a legitimate one. This approach applies to both phishing and non-phishing website classifications and will have the potential to boost the efficacy of current phishing detection systems. The limited dataset that was utilized in the project affects the prediction and efficiency of the model. Therefore, the potential for improvement in the approach includes adding more datasets for training, testing, and using different models in making predictions. The project highlights the need to invest in detecting and preventing phishing attacks, as they pose a serious threat to online security.

Arathi Krishna V et al. (2022) investigated the manner in which to categorize malicious URLs within a dataset containing both benign and phishing URLs [2]. They explored dataset randomization, feature engineering, and feature extraction methods, including lexical analysis, host-based features, and statistical analysis. Comparative analysis using various classifiers consistently yielded improved accuracy with dataset randomization. The authors used a straightforward approach involving regular expressions for the extraction of features and suggested the potential for further experimentation to enhance system accuracy. They emphasized the significance of ongoing training with updated data to enhance model performance and pointed out that the dataset may have been out of date. Due to the short lifespan and inaccessibility of phishing websites, the study did not use content-based features. This project lacks the integration of classification-based lexical analysis and rule-based URL content analysis for a comprehensive phishing URL identification solution.

According to Adarsh Mandadi et al. (2022), a model is developed by using an ML algorithm, that's employed to identify phishing URLs and notify the user in advance [3]. The parameters of the URL are extracted, which are entered by the end user in the respective field, and this acts as input data for the machine learning model. The model processes this and gives the output as to whether it seems to be phishing or genuine. The algorithms that are employed in developing this model are Random Forest and Decision Tree. From the confusion matrix obtained, it’s mentioned that a random forest has a higher accuracy than a decision tree, so this Random Forest algorithm is stored with the use of a pickle for deployment. The model is deployed as a webpage with the help of Flask API. This webpage, created using HTML (Hyper Text Markup Language), includes a textbox and a submit button. Whenever a user enters a URL and clicks on the submit button, this URL will be processed through the model and return a value as a binary that is 0 or 1. If the returned value is '0', then the output is displayed as "Legitimate" although if the returned value is other than ‘0’, the output is displayed as "Phishing". After training the model, the accuracy obtained was 87.0% for RF, and the reliability of the Decision tree was 82.4%.

Zhixin Zhou et al. (2022) [4] highlighted that website detection based on blacklist or webpage characteristics cannot meet the time and timeliness requirements of batch phishing detection. To get around this obstacle, they created and verified a dual-weight random forest algorithm specifically tailored for phishing website detection. Experimental results demonstrated that representative features could be effectively identified through a clustering algorithm. These features were subsequently put to use to generate a decision tree, resulting in improved validity of the detection model as well as a reduction in missed detections. The average accuracy of system detection is 96.24%, the average false positive rate is 14.05%, and the average false negative rate is 85.95%. The authors stated that the next phase of research will focus on optimizing the algorithm's complexity, increasing its efficiency, and decreasing the overall time consumption of the detection process.

Kamal Omari (2023) studied the critical area of phishing website detection, examining various machine-learning approaches as well as their efficacy against the growing cybersecurity threat of phishing attacks [5]. The growing prevalence of these fraudulent activities has made it difficult for individuals and organizations worldwide, emphasizing the need for robust and efficient detection methods to combat these deceptive websites. Through a thorough examination of their research findings and comparison with other relevant studies, the authors discovered promising findings about the effectiveness of several machine-learning techniques in identifying phishing attacks. Among the seven ML models, Logistic Regression (LR), k-nearest neighbors (KNN), Support Vector Machine (SVM), Naive Bayes (NB), Decision Tree (DT), Random Forest (RF), and Gradient Boost, the GB and RF models outperformed the others, with accuracy rates comparable to those stated in the literature. These models are strong contenders for practical use in detecting phishing, significantly improving cybersecurity, and safeguarding users from the dangers of phishing attacks.

Chong Kim Ying et al. (2021) have implemented the detection of phishing websites through machine learning (ML) [6]. According to the findings and discussions of each model developed, the researcher successfully achieved the main goal of this study, which is to identify and classify websites or URLs as legitimate or phishing using machine learning. Various methods were employed to detect phishing websites. The researcher opted to employ and evaluate datasets of phishing websites containing 11,055 observations across 15 variables sourced from Kaggle.com, with the "result" variable designated as the independent variable. For the experiments, the implemented algorithms used for machine learning included Decision Tree, KNN, and Random Forest. The motivation for choosing these methods for machine learning was to investigate lazy and eager learners' classification model construction processes. Even though KNN achieved the highest accuracy, it is highly susceptible to overfitting. RF is regarded as superior to the other two algorithms, which is supported by the assertion that the RF algorithm is the most precise and appropriate, with the lowest false-negative rate.

Phishing is a serious cybersecurity threat, so accurate and efficient anti phishing measures are critical for protecting online users and organizations. Arathi Krishna V et al. (2022) investigated dataset randomization and feature engineering, suggesting potential improvements in accuracy through further testing. Adarsh Mandadi et al. (2022) built a model utilizing RF and Decision Tree algorithms, achieving an 87.0% accuracy rate with Random Forest. Zhixin Zhou et al. (2022) developed a specialized dual-weight random forest algorithm, reaching an average accuracy of 96.24%. Kamal Omari (2023) emphasized the effectiveness of Gradient Boost and RF models in phishing detection, matching the performance levels reported in prior research. Chong Kim Ying et al. (2021) employed a range of machine learning techniques, with Random Forest showing the best accuracy and the fewest false negatives. Together, these studies highlight the importance of the use of machine learning for enhancing phishing detection and stress the importance of ongoing model training and dataset improvements to achieve optimal cybersecurity outcomes.

V. PROPOSED APPROACH

A proposed system is a conceptual solution that aims to accomplish specific objectives or solve specific problems. It pertains to intended modifications or adjustments meant to implement a brand-new system or enhance the operation of an already existing one. The goal, key components, architectural design, chosen technologies, UI considerations, integration techniques, security protocols, scalability, and performance requirements are typically included in defining a proposed system. It also addresses protocols for testing, resource requirements, cost estimates, documentation strategies, and implementation plans.

The proposed approach to our project mainly focuses on 4 modules, each interconnected with the other modules, thus binding all the features together. The modules are the Driver module, Web Module, AI Prediction Module, and Feature Extraction from URL module.

App module

This is the main module, which serves as a driver and connecting module between all the other modules. This module does the function of accepting the URL or interaction with the user and then routing data views accordingly after prediction and rendering of results, thus enabling connectivity with the other modules to work with.

Web module

The Web Module is the display module, which views and interfaces with the user. This module is rendered based on the results from the prediction module. This module allows the user to upload a custom data set, enter the URL to scan, and contains the view that's going to be displayed on the screen after the AI model predicts the results.

AI prediction module

This module is the heart of the project; it is the classification. model built around a random forest classification algorithm. This module accepts the results from the feature extraction module as features of the URL and does the prediction using the data set of features provided for the random forest algorithm to learn from and do the prediction. The outcomes

of this are sent to the app module, which renders the view to be presented with the assistance of the web module.

Feature extraction module

As the URL itself cannot be consumed by the AI model, this module helps the AI model work by providing appropriate data input for features extracted from the URL which the AI module can then proceed to predict.

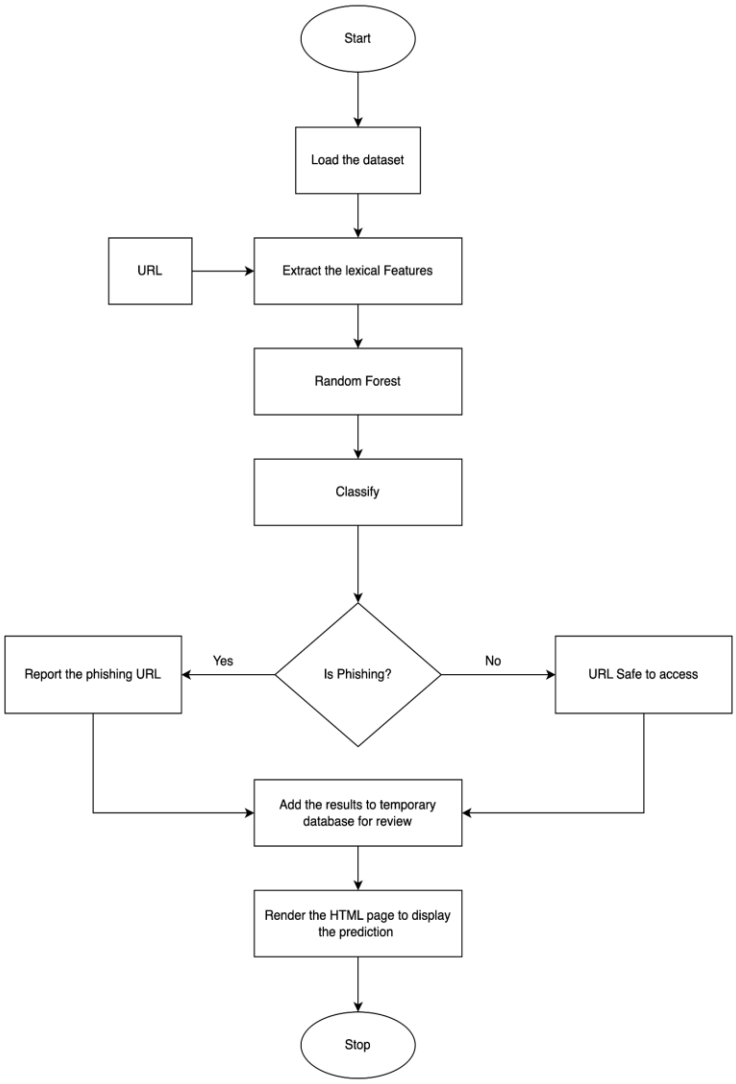


Figure 1: Architecture of the system

Data Set Description

The dataset contains 11000+ website URLs, of which 6157 are legitimate and 4897 are phishing.

The features of the set of data were determined using the Python programming language. This model is trained using the following 30 parameters extracted from websites:

'Index', 'UsingIP', 'LongURL', 'ShortURL', 'Symbol@', 'Redirecting//', 'PrefixSuffix-', 'SubDomains', 'HTTPS', 'DomainRegLen', 'Favicon', 'NonStdPort', 'InfoEmail', 'HTTPSDomainURL', 'RequestURL', 'AnchorURL', 'LinksInScriptTags', 'ServerFormHandler', 'AbnormalURL', 'WebsiteForwarding', 'StatusBarCust', 'DisableRightClick', 'UsingPopupWindow', 'PageRank', 'IframeRedirection', 'AgeofDomain', 'DNSRecording', 'WebsiteTraffic', 'GoogleIndex', 'LinksPointingToPage', 'StatsReport'

Methodology

In the proposed approach, we have made use of an RF classifier to build our AI prediction model. Random Forest relies on ensemble learning, a technique that addresses complex problems by blending multiple classifiers to enhance model performance. This classifier employs multiple decision trees on various data subsets and averages their predictions to boost the overall reliability of the data set. Instead of relying on a single decision tree, a random forest collects predictions from all the trees and determines the outcome using the majority of these predictions. The precision of a random forest increases with the variety of trees it contains, reducing the risk of overfitting. The Random Forest model that has been trained classifies a given URL as legitimate or phishing according to the learned patterns derived from the training information. By aggregating the predictions from numerous decision trees, RF provides more robust predictions for classifying URLs as either genuine or phishing.

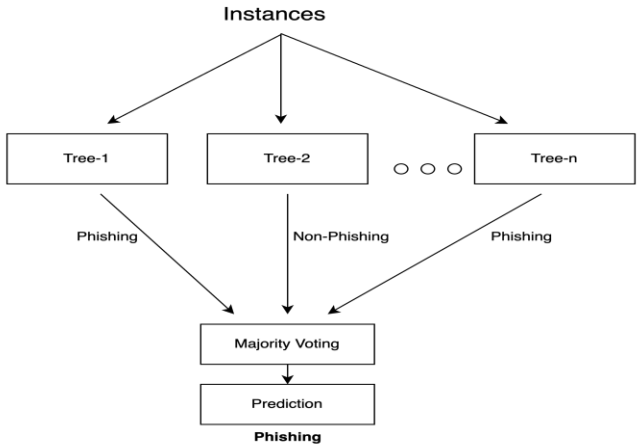


Figure 2: Classification of the URLs using RF

OUTPUT:

The objective of this methodology is to efficiently extract and analyze the features from the website URLs, which are then passed to the AI model built using the RF algorithm. The RF classifier provides a far more reliable and accurate prediction of web addresses by aggregating the predictions of multiple decision trees. This will help us get more accurate predictions and prevent us from visiting phishing websites.

V. IMPLEMENTATION AND RESULT

The dataset with extracted features is divided into testing and training sets using a 70-30 ratio. The data used for training is utilized to teach the RF, while the testing data is employed to evaluate and find the success rate of the algorithm.

	precision	recall	f1-score	support
-1	0.96	0.97	0.97	1436
1	0.98	0.97	0.97	1881
accuracy			0.97	3317
macro avg	0.97	0.97	0.97	3317
weighted avg	0.97	0.97	0.97	3317

Figure 3: RF Metrics

The above figure represents the classification metrics of the RF algorithm, which is employed to find how precise the model is. The algorithm has done a great job with high precision, recall, and F1-score values for the two classes (-1 and 1). The high F1-score indicates a good balance between precision and recall, which is crucial when the classes are imbalanced.

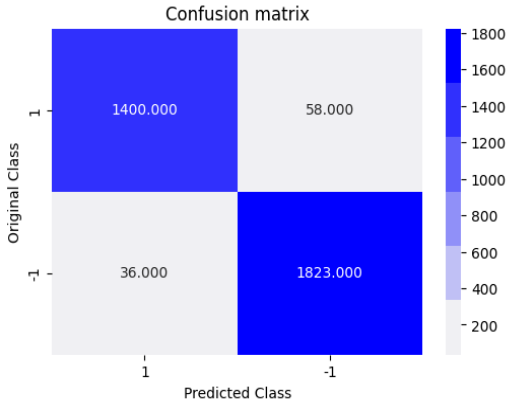


Figure 4: Confusion matrix

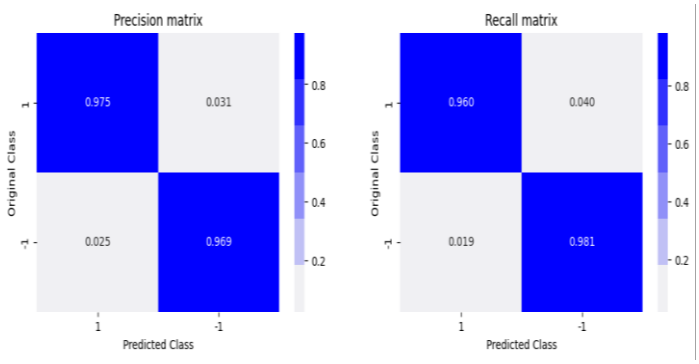


Figure 5: Precision and Recall Matrix

The classification model performs admirably, achieving an overall accuracy of 97%. Both classes (-1 and 1) have high precision (96% and 98%, respectively) and recall (97% for each). This suggests that the model can distinguish and predict both target classes with nearly equal accuracy.

VI. MODEL DEPLOYMENT

From the above classification report and the confusion matrix, it is clearly shown that a random forest has a high accuracy. The model is deployed as a webpage with the help of Flask API. This webpage contains a textbox and a submit button which is developed using Hyper Text Markup Language (HTML) and Cascading Style Sheet (CSS).

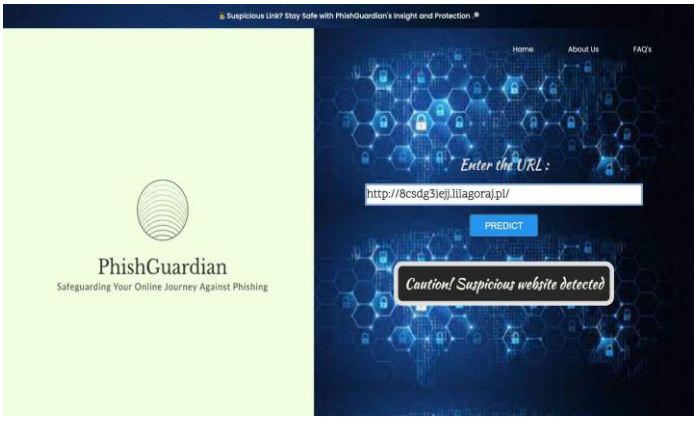
Whenever we enter a URL and click on a predict button then this URL will be processed by the model and returns a value as a binary that is 1 or -1. If the returned value is '1' then the output is displayed as "The URL seems to be safe to view." and if the returned value is '-1' the output is displayed as "Caution! Suspicious website detected".

Figure 6: PhishGuardian's Home Page

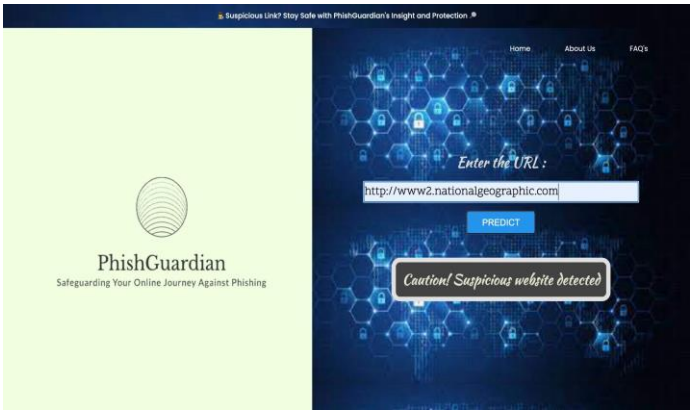
Prediction of phishing URLs

Here we have passed the phishing website URL to get the correct prediction.

URL-1: http://8csdg3iejj.lilagoraj.pl/

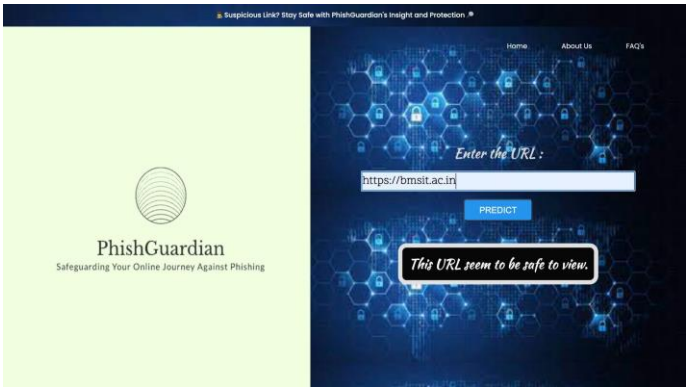


URL-2: http://www2.nationalgeographic.com

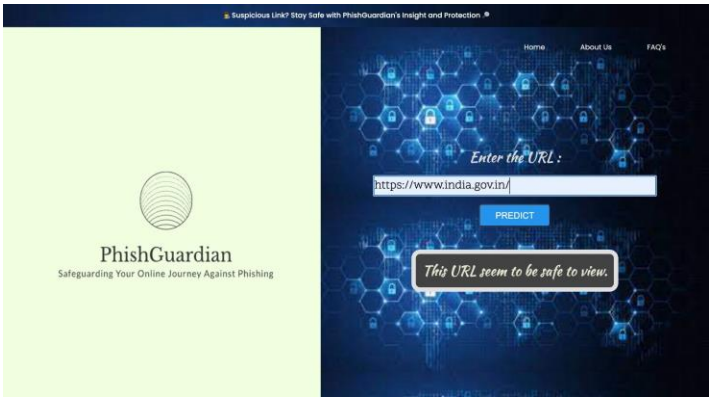


Prediction of legitimate URLs

URL-1: https://bmsit.ac.in



URL-2: https://www.india.gov.in/



VII. CONCLUSIONS

In conclusion, we explored the efficacy of the RF machine learning algorithm in identifying phishing websites. The parameters of the URL are extracted, which are entered into the user's browser in the respective field, and this acts as input data for the prediction model. The model processes this and gives the output as to whether it is phishing or legitimate. Using a dataset with extracted features, we trained RF machine learning models and achieved an impressive 97% success rate in terms of accuracy. The model's strong performance is evident from the classification report, which showed a precision of 0.96 and 0.98, recall of 0.97 and 0.97, and an F1-score of 0.97 for the benign and phishing categories, respectively. These results confirm the Random Forest algorithm's ability to differentiate between phishing and benign URLs, highlighting its advantages compared to traditional machine learning approaches. The elevated precision and recall metrics underscore the model's precision in accurately flagging malicious websites.

Our research illustrates that the RF algorithm is an effective tool in the fight against phishing attacks. It provides a dependable and efficient method for identifying malicious websites, affirming its strength and capacity to bolster cybersecurity measures. By shielding online users from the dangers of phishing activities, the Random Forest algorithm plays an essential role in strengthening cybersecurity safeguards.

REFERENCES

[1] APWGG, “Phishing Activity Trends Report,” 2023.

[2] Arathi Krishna V, Anusree A, “Phishing Detection using Machine Learning based URL Analysis: A Survey”, 2022 IJRTI, Volume 7, Issue 7

[3] Adarsh Mandadi, Saikiran Boppana, Vishnu Ravella, R Kavitha, “Phishing Website Detection Using Machine Learning”, 2022 IEEE 7th International Conference for Convergence in Technology (I2CT), pp. 1-4, 2022.

[4] Zhixin Zhou, Chenghaoyue Zhang, “Phishing website identification based on double weight random forest”, 2022 IEEE 3rd International Conference on Computer Vision, Image and Deep Learning & International Conference on Computer Engineering and Applications (CVIDL & ICCEA)

[5] Kamal Omari, “Comparative Study of Machine Learning Algorithms for Phishing Website Detection”, 2023 International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 14, No. 9, 2023

[6] Chong Kim Ying, Farashazillah Yahya, Ryan Isaac W Mahibol, Magnus Bin Anai, Sidney Allister Frankie, Eric Ling Nin Wei, Rio Guntur Utomo, “Detection of Phishing Websites using Machine Learning Approaches”, 2021 International Conference on Data Science and Its Applications (ICoDSA)