Phishing Website Detection

B.Raga Madhuri[1],Tejaswini Vaddi[2],Priyanka kalidindi[3],N.devika[4]
Computer Science and Business System,Department of Information Technology,
S.R.K.R Engineering College(A) [1,2,3]
Assistant Professor,Department of Information Technology,
S.R.K.R Engineering College(A),Bhimavaram,A.P[4]
bommiraga@gmail.com[1], tejaswinivaddi31@gmail.com[2],
priyankakalidindi3714@gmail.com[3]

ABSTRACT: *Phishing websites pose a significant trouble to cybersecurity, serving as the entry point for multitudinous attacks risking the confidentiality, integrity, and vacuity of sensitive data for both companies and consumers. Despite decades of exploration, the automatic discovery of phishing websites remains grueling due to the evolving nature of similar attacks. Being slice-edge results frequently bear expansive homemade point engineering and struggle to identify new phishing attempts effectively. Hence, there's a pressing need to develop strategies capable of fleetly detecting and mollifying zero-day phishing attacks. In this study, we propose a new approach to automatically describe phishing websites by using machine literacy ways for point birth and analysis. Through a comprehensive review of being literature, we identify crucial features present in phishing web runners that can be employed to discern vicious intent. Our system aims to address the limitations of former approaches while enhancing the effectiveness and delicacy of phishing discovery. We collect a different dataset comprising known phishing websites and licit bones to train machine literacy models for prognosticating the liability of a website being involved in phishing conditioning. By employing advanced machine learning algorithms, we seek to ameliorate the discovery capabilities of our system and effectively identify arising phishing pitfalls. The primary idea of this exploration is to develop a robust and scalable result for detecting phishing websites, thereby enhancing the cybersecurity posture of associations and individuals. Through rigorous trial and evaluation, we aim to demonstrate the effectiveness of our approach in anticipating and combating phishing attacks in real-world scripts. This exploration focuses on strategies for detecting phishing attacks.*

Keywords: *Phishing,Machine learning algorithms,Feature extraction,Cyber-attacks, Uniform resource locators*

I. Introduction

Phishing has emerged as the biggest issue, affecting people, businesses, and even entire nations. 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, a vast quantity of data is continuously downloaded and uploaded to the Internet. Social engineering techniques use spoof emails purporting to be from reliable companies and organizations to send visitors to phony websites that trick them into divulging personal information like usernames and passwords. Technical tactics typically entail installing malicious software on computers to directly steal credentials. These methods are often used to intercept usernames and passwords for internet accounts. Historically, heuristics or static rule-based systems have been used for phishing detection.

Machine learning's contribution to improving phishing detection skills: The development of increasingly complex and precise detection models is made possible by machine learning, which is a critical factor in improving phishing detection capabilities. Machine learning is essential in this situation for the following reasons:

Extraction of features: Algorithms for machine learning can automatically extract pertinent information from webpages, phishing emails, and network traffic. To mention a few of these aspects

are email header data, content analysis, URL structure, and user behavior.

Patterns.Classification and Prediction: Machine learning models, including decision trees, support vector machines, and neural networks, can be trained on label datasets to classify emails, URLs, or network traffic as either phishing or legitimate.

Adaptability and Scalability: Machine learning algorithms can evolve and progress over time by continuously learning from new data. Because of their adaptability, detection models can be effective even in the face of newly identified threats and evolving phishing techniques. Furthermore, machine learning techniques may be expanded to efficiently handle large volumes of data, enabling the analysis of massive email databases or network traffic logs.

Ensemble Learning: To increase detection accuracy, ensemble learning techniques combine many machine learning models. Examples of these techniques are gradient boosting and random forests. The utilization of individual models' uniqueness enables ensemble approaches to proficiently manage intricate and varied phishing assault scenarios.
All things considered, machine learning greatly improves the ability to detect phishing attempts by facilitating the creation of complex detection models that can adjust to changing threats and accurately pinpoint fraudulent activity in a variety of network traffic and communication formats.

II. LITERATURE SURVEY

1. Detecting Phishing Websites Using Machine Learning(2019) The study focuses on the categorization of websites using phishing techniques. This study looks at the effectiveness of machine learning techniques for identifying phishing sites by examining website attributes. In order to better understand the limitations and capabilities of these algorithms, it is important to know how effectively they work.

2. Machine Learning Algorithms Evaluation for Phishing URL Classification(2021) Machine Learning (ML) has emerged as an effective approach for detecting and classifying these phishing URLs. The detection accuracy of each algorithm was computed, and URL features were extracted using lexical analysis.

3. Phishing Attack Detection on Text Messages Using Machine Learning Techniques(2022)The use of machine learning techniques has expanded the range of phishing website detection options. It uses machine learning (ML) approaches to identify phished messages, such as KNN, Support Vector Classification, Random Forest Classifier, and Naive Bayes' Classifier.

4. The Emergence Threat of Phishing Attacks and The Detection Techniques Using Machine Learning Models(2021) The primary goal of this study is to evaluate the performance of various algorithms, including Decision Trees, Random Forest, and K-Nearest Neighbors. It has been discovered that some of these algorithms require a lot of training time and have false negative rates.

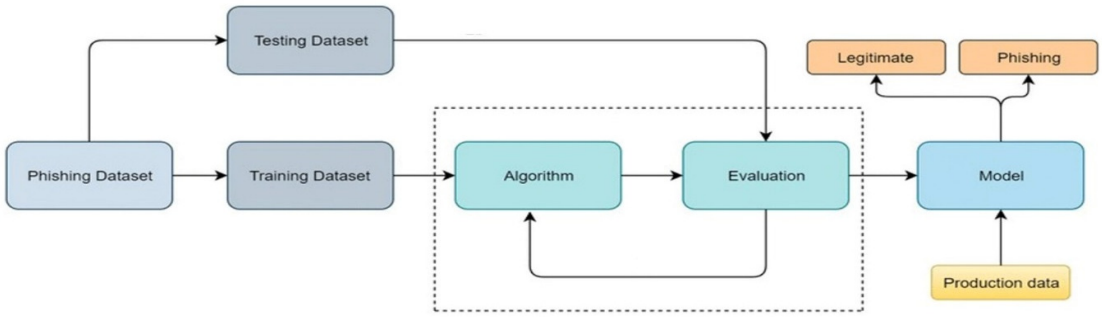
III. PROBLEM STATEMENT

The formulation of the problem emphasizes the need for an improved approach that can identify phishing websites with more accuracy and adaptability. Programmers' constantly changing techniques often provide a barrier for traditional phishing detection tools, making it challenging to distinguish between legitimate and fraudulent websites. This increases the difficulty of safeguarding user's sensitive data from online threats.

The aim is to develop a framework that utilizes machine learning algorithms that have been trained on massive datasets to identify evolving trends and features in phishing site designs. By doing this, the method aims to stay up to date with threats that aren't being employed and provide strong and reliable

discovery capabilities. The problem statement emphasizes the necessity of managing certain deterrents in phishing locations to boost online client security.

IV. ARCHITECTURE



ALGORITHMS:

1. Classifier: Gradient Boosting: Type: Ensemble Learning Learning-Features: Combines the predictions from many decision trees to improve predictive performance. As a member of the boosting algorithm family, it adds predictors to an ensemble successively, correcting the previous one in the process.
2. Random Forest: Type: Ensemble Learning; Features: Combines many decision trees to yield more accurate predictions. They don't require scalable data, are typically highly powerful, and operate well with little parameter tweaking.
3. Multi-layer Perceptron: Type: Neural Network- Features: MLPs are like interconnected layers of neurons in the brain, used for complex tasks like recognizing patterns in images and speech.
4. Support Vector Machine: Type: Discriminative - Features: Determines the ideal line to divide points on a graph, or the best hyperplane to divide data into classes. SVM is effective for both regression and classification problems, particularly when dealing with high-dimensional data.
5. Classifier XGBoost: Gradient boosting is the type. Features: Its performance, scalability, and speed have made it a popular option for a variety of machine learning applications in both academia and industry.

V. IMPLEMENTATION

Input:
- Dataset of phishing URLs and features
- Machine learning models (e.g., Gradient Boosting Classifier, Random Forest, etc.)

Output:
- Trained model for phishing website detection

Steps:
1. Data Collection: Collect phishing URL datasets from sources like Phish Tank, ensuring the inclusion of URL characteristics and domain attributes.

2. Data Preprocessing: Load dataset with Pandas, then handle missing values, outliers, and anomalies, followed by exploration for structural understanding and distributions.

3. Data Splitting With the dataset at hand, the data was strategically divided into two distinct subsets: the training set and the test set. This partition is crucial to evaluate the performance of machine learning algorithms accurately.

4. Feature Extraction: Extract domain-based, HTML-based, and address-based features such as domain, IP address, URL length, depth, and redirection; ensure categorical feature handling and encoding as needed.

5. Model Training: Train machine learning models like Gradient Boosting Classifier, Random Forest, Support vector machine, etc on split training and testing datasets (e.g., 80% for training, 20% for testing), evaluate using metrics such as accuracy, precision, recall, F1-score, and tune hyperparameters for performance enhancement.

6. Model Selection: Choose the top-performing model based on evaluation metrics such as accuracy, precision, recall, and F1-score from the testing dataset.

7. Deployment:Integration with Flask Framework In order to ensure a smooth operation, the Flask framework was used. Not only does Flask make it easier to create a web interface, The tool was easy to use, and users eventually received accurate predictions for phishing websites.

VI. VISULAISATION OF DATA

	ML Model	Accuracy	f1_score	Recall	Precision
0	Gradient Boosting Classifier	0.974	0.977	0.994	0.986
1	Multi-layer Perceptron	0.970	0.974	0.994	0.980
2	Random Forest	0.966	0.970	0.993	0.989
3	Support Vector Machine	0.964	0.968	0.980	0.965
4	XGBoost Classifier	0.549	0.549	0.554	0.554

FIGURE 1. Comparison of Machine Learning Models

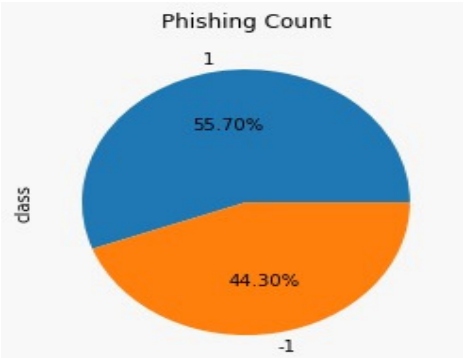


Figure 2. Phishing Count

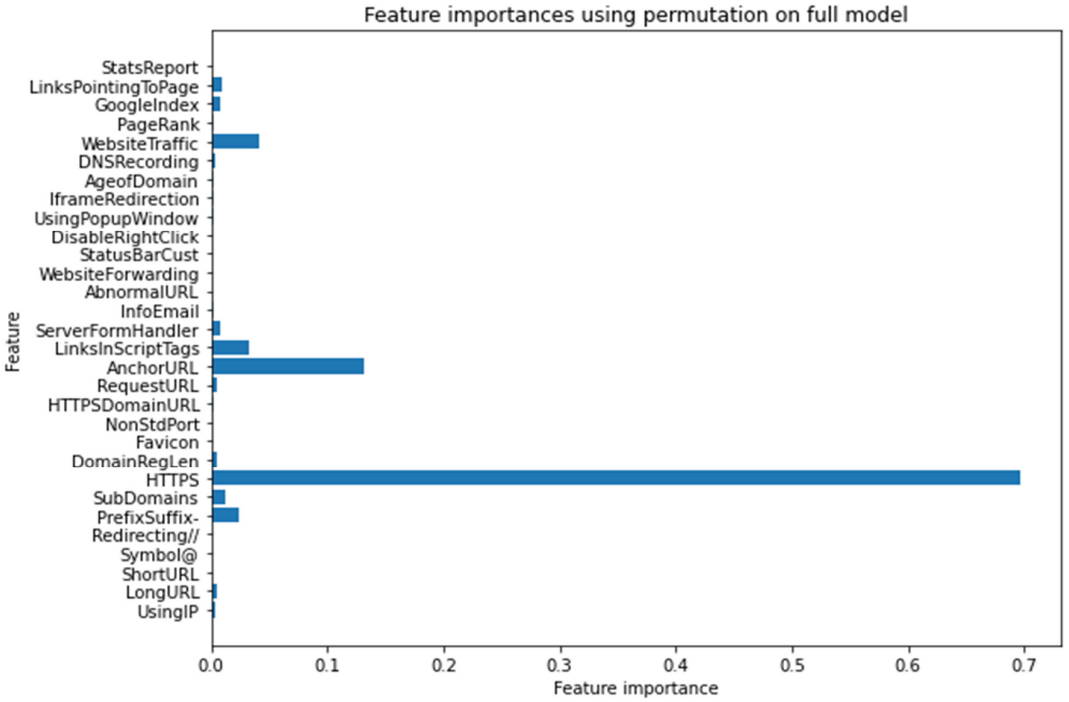


Figure 3. Feature Extraction

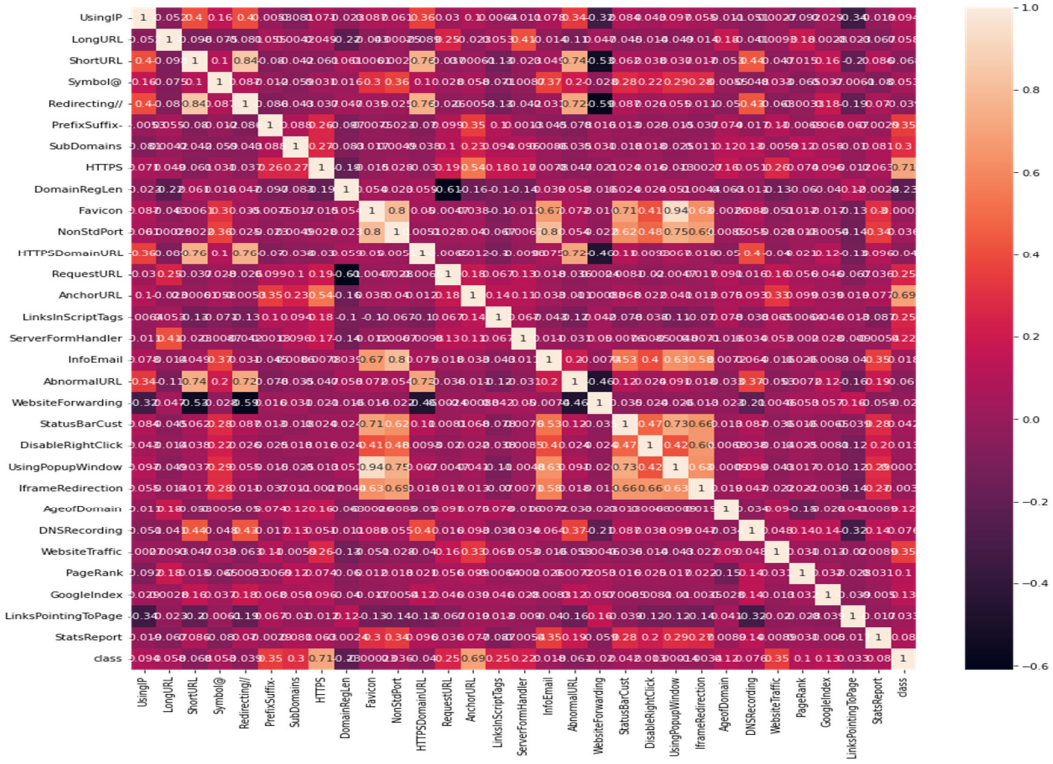


Figure 4. Heat Map

VII. CONCLUSION

Phishing demonstrations are becoming a sophisticated threat to this rapidly expanding world of technology. Every country is concentrating on cashless transactions, online commerce, paperless ticketing, and other initiatives to keep up with the rapidly expanding global community. However, phishing is beginning to impede this progress. People no longer believe that the internet is reliable. AI has the potential to be used to gather data and compile remarkable information pieces. A layperson who has no idea how to spot a security risk would never put themselves in danger by transacting money over the internet. Phishers target the installment market because of the cloud's greatest advantages. The research aims to explore this area by presenting a case study of using machine learning to identify phishing websites. Its goal was to create a phishing detection system that is accurate, efficient, and economical by utilizing machine learning tools and methodologies. The Python project was created and executed within the Anaconda IDE. To do this, the suggested approach employed four machine learning classifiers, and a comparison of the four algorithms was conducted. Additionally, a high accuracy score was attained. The eight methods that are employed are the following: XGBoost Classifier, Logistic Regression, Decision Tree, K-Nearest Neighbors, Random Forest, Multi-layer Perceptron, Support Vector Machine, Gradient Boosting Classifier, and Naive Bayes Classifier. While all eight classifiers produced encouraging results, the Gradient Boosting Classifier performed the best, scoring an accuracy of 97.4%. When utilizing different datasets and algorithms, the accuracy score may change and may be higher or lower than that of the Gradient Boosting Classifier. Because the Gradient Boosting Classifier is an ensemble classifier, its accuracy is very high. To determine whether a URL is phishing or authentic, this model can be used in real-time.

VIII. REFERENCES

[1] V. K. Nadar, B. Patel, V. Dev mane and U. Bhave, "Detection of Phishing Websites Using Machine Learning Approach," 2021 2nd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2021, pp. 1-8, doy: 10.1109/GCAT52182.2021.9587682.

[2] Mandadi, S. Boppana, V. Ravella and R. Kavitha, "Phishing Website Detection Using Machine Learning," 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India, 2022, pp. 1-4, doy: 10.1109/I2CT54291.2022.9824801.

[3] 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), Greater Noida, India, 2021, pp. 78-82, doy: 10.1109/ICACITE51222.2021.9404714.

[4] M. A. Ivanov, B. V. kulchinicova, I.V. chugunkov and A. M. Plaksina, "Phishing Attacks and Protection Against Them," 2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus), St. Petersburg, Moscow, Russia, 2021, pp. 425-428, doi: 10.1109/ElConRus51938.2021.9396693.

[5] M. Kathiravan, V. Rajasekar, S. J. Parvez, V. S. Durga, M. Meenakshi and S. Gow salya , "Detecting Phishing Websites using Machine Learning Algorithm," *2023 7th International Conference on Computing Methodologies and Communication (ICCMC)*, Erode, India, 2023, pp. 270-275, doy: 10.1109/ICCMC56507.2023.10083999