Detection of Phishing Website Using Gradient Boosting Algorithm

Yawalkar Prasad Pramod[1], Dr. Prashant N. Chatur[2], Dr. Kamlesh A. Waghmare[3]

[1]*Student, Department of Computer Science and Engineering, Government College of Engineering, Amravati, Maharashtra, India*
[2]*Professor, Department of Computer Science and Engineering, Government College of Engineering, Amravati, Maharashtra, India*
[3]*Assistant Professor, Department of Computer Science and Engineering, Government College of Engineering, Amravati, Maharashtra, India*

ABSTRACT

Phishing attacks have emerged as a significant risk to online security, frequently deceiving users by impersonating trustworthy websites to obtain sensitive data. Traditional detection methods often struggle to keep up with the ever-evolving tactics employed by phishers. This paper proposes a fast and effective machine learning-based method for identifying phishing websites, using only the URLs as input features. The suggested model identifies and examines a diverse set of lexical and statistical features from URLs, including domain length, the presence of special characters, and specific token patterns, all of which are indicative of phishing activities. Gradient boosting, renowned for its exceptional predictive accuracy and capacity to handle intricate datasets, is utilized to construct a reliable classifier. The model is trained and tested using publicly available phishing datasets and performs exceptionally well, achieving high accuracy, precision, and recall compared to other traditional classifiers. The result showcase the effectiveness of url feature-based detection in identifying phishing websites, providing a lightweight solution for improving cybersecurity measures.

Keyword : -, *Phishing Attack, Machine Learning, URL, Cybercrime, Personal Information.*

1. INTRODUCTION

In the modern era of technology, phishing attacks have emerged as a major concern for online security, as attackers employ advanced methods to trick users into divulging confidential data due to that cybercrime cases increasing day by day.

The widespread adoption of the internet has not only made our lives more convenient but has also exposed us to various security risks. One of the most common cyber threats in today's digital landscape is phishing, where malicious actors deceive users into divulging confidential information by impersonating legitimate organizations through fraudulent websites. The Anti-Phishing Working Group (APWG) reports that phishing attacks have experienced a significant rise in recent years, with millions of users worldwide falling victim to these scams[1].

Phishing is a type of online identity theft where an attacker sends deceptive emails and creates fake websites to deceive unsuspecting customers into revealing sensitive information like bank account details, website login credentials, and more[2]. Phishing websites are expertly designed to deceive users into thinking they are engaging with a legitimate organization, ultimately enabling unauthorized access to their personal information. Traditional methods of blocking phishing attempts, such as rule-based and blacklist approaches, have been against the rapidly evolving and advanced tactics employed by phishers[3]. Phishing websites, in particular, are highly dangerous, as they can closely resemble legitimate sites, making it difficult for users to differentiate between trustworthy and fraudulent websites. To tackle this issue, machine learning techniques have been gaining popularity for automatically detecting phishing websites by analyzing features extracted from URLs, domain metadata, and website content[4].

One promising approach is to leverage the features of URLs to identify phishing websites. URLs (uniform resource locators) hold significant information regarding a website's identity, structure, and content. By examining various characteristics of a website's URL, such as its length, sentence structure, and keyword usage, machine

learning algorithms can develop the ability to identify patterns associated with phishing websites. This method provides several benefits, such as immediate identification and low computational requirements.

The gradient boosting algorithm is an appropriate machine learning technique for identifying phishing websites by analyzing their URL features. This algorithm is particularly adept at dealing with intricate datasets and can successfully identify non-linear connections between different features. By repeatedly combining several weak models, gradient boosting can create a strong and precise predictive model. Its capability to handle complex data and resistance to overfitting make it an appealing option for this task. This method entails identifying important features from a dataset of websites that have been labeled as either phishing or legitimate, and then using a gradient boosting model to classify new, unknown urls as either phishing or legitimate. By combining the power of gradient boosting and url feature analysis, this method can offer a reliable solution for identifying phishing websites and safeguarding users from online scam[5].

2. PROBLEM STATEMENT

Phishing attacks have become a major concern in the digital realm, where fraudulent websites masquerade as genuine ones to trick users into divulging personal information like usernames, passwords, and financial data. Traditional security measures such as blacklists and rule-based detection systems frequently fall short in keeping pace with the constantly evolving tactics employed by cybercriminals. Consequently, there is an need for intelligent, data-driven methods that can effectively and accurately identify phishing websites[6].

3. OBJECTIVES

The objectives of the system are as follows :-

1. Data Collection : Collect a dataset that includes various features. For this system, we have collected dataset from Kaggle.
2. Feature Extraction : Identify and Extract features such as:
    - Length, Presence of special characters etc are URL-based features.
    - Age of domain, DNS record etc are Domain-based characteristics.
    - The presence of forms, iframes, and pop-ups etc are content-based features.
    - Certificate Authority, HTTPS usage etc are HTTPS/SSL - related features
3. Model Development : Utilize the gradient boosting algorithm to train a classification model.
4. Model Evaluation : Employ Metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC to assess the performance of the model.

4. METHODOLOGY

Machine learning models are highly valuable assets in the field of phishing detection, providing the capability to automatically detect and prevent phishing URLs, even those cleverly designed to evade traditional detection methods.

Fig 1, depicts the complex mechanisms involved in a machine learning model specifically designed to detect phishing attempts. Its training process relies heavily on a large dataset that includes both known phishing websites and legitimate websites. After being extensively trained, this model is now ready to promptly classify new web addresses as either authentic or fraudulent.

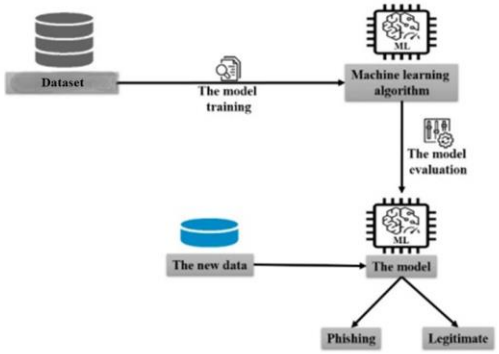

Fig. 1. Machine Learning Model for Phishing URL Detection

Our main goal in this study was to evaluate the effectiveness of the gradient boosting algorithm in accurately identifying phishing websites within a specific dataset. We specifically chose the gradient boosting algorithm because it has a highest accuracy rate compare to other machine learning classifier.

We partitioned our data, assigning 80% for training the model and 20% for evaluating its performance. By following this systematic approach, we thoroughly assessed the machine learning algorithms using well-known metrics such as Accuracy, Precision, Recall, F1-Score, and ROC-AUC. These metrics provided valuable insights into how well our models performed, offering a comprehensive understanding of their efficacy in identifying phishing websites within the dataset. This examination allowed us to recognize the strengths and weaknesses of the gradient boosting algorithm in this particular scenario.

Fig 2, represents the model's path way through the evaluation process, providing a comprehensive overview of the steps involved.

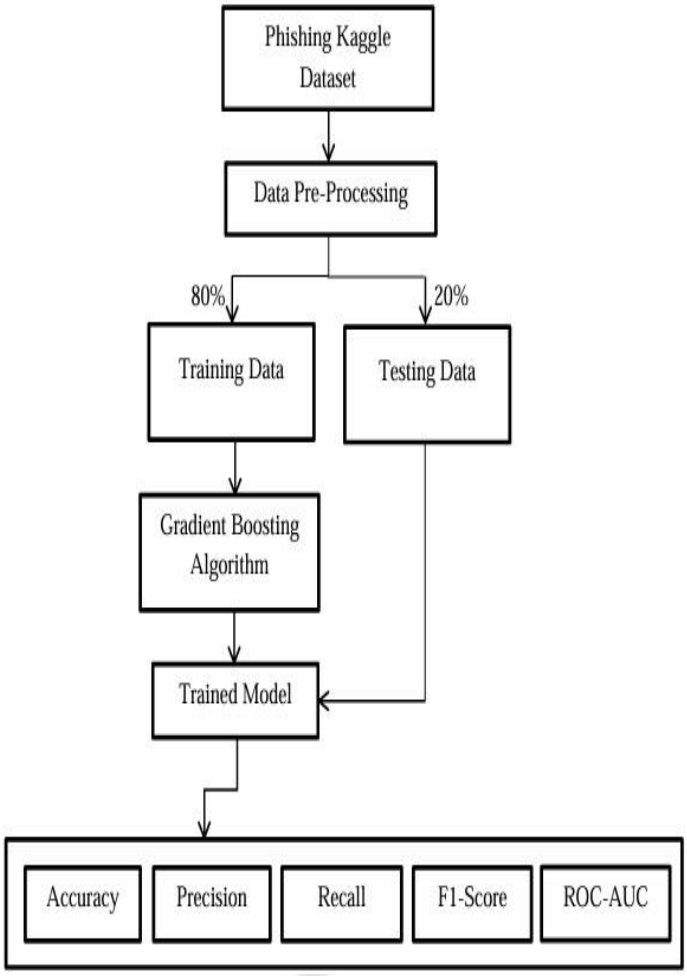

Fig 2. Model's Flowchart

4.1 The Dataset

A standard dataset of phishing attacks from Kaggle was used for machine learning processing which included 11,055 records. Each entry comprised 32 unique features.

4.2 The Gradient Boosting

A gradient boosting classifier is a highly effective machine learning algorithm that is frequently employed for identifying phishing websites. It belongs to ensemble learning family and is frequently utilized to enhance the accuracy and resilience of binary classification tasks, such as differentiating between genuine and fraudulent websites.

Gradient boosting is a machine learning method that combines several weak learners (usually decision trees) to build a robust predictive model. The main objective is to continuously add decision trees, with each new

tree rectifying the mistakes made by the previous ones. This process is guided by the gradient of a loss function, which measures the disparity between the predicted and actual values[7].

5. RESULTS AND DISCUSSION

The outcomes derived from the experiments conducted with the help of the scikit-learn tool demonstrate the efficiency of the Gradient Boosting Model in identifying phishing domains (Table 1). With an accuracy of 0.974, the model showcases a high level of correctness in its predictions. The f1_score of 0.977 indicates a balanced performance in precision and recall, reflecting the model's ability to capture both true positives and minimize false negatives. The recall value of 0.994 highlighted model's ability to accurately identify the majority of actual phishing instances. Furthermore, The precision of 0.986 indicates a low false positive rate, underscoring the model's efficiency in avoiding incorrect predictions of legitimate websites as phishing sites. The ROC-AUC score of 0.996 further reinforces the model's ability to accurately distinguish between positive and negative instances was tested across different thresholds.

Table 1. The Evaluation Results

Classifier	Accuracy	F1 Score	Recall	Precision	ROC-AUC
Gradient Boosting	0.974	0.977	0.994	0.986	0.996

Our model's performance in accurately identifying phishing websites was significantly better, with higher F1 score, ROC-AUC Score, and overall accuracy. This comparison highlights the strength and effectiveness of our gradient boosting approach in identifying phishing websites.

6. CONCLUSION

This research aimed to investigate the identification of phishing websites using a machine learning approach, with a particular focus on the gradient boosting algorithm and feature extraction methods. By examining various features, including URL characteristics, domain identity, and webpage behavior, we were able to train a reliable model that could effectively differentiate between genuine and phishing websites.

The gradient boosting algorithm demonstrated exceptional performance because it could merge multiple weak learners into a robust predictive model, enhancing accuracy through iterative learning. The outcomes showcased exceptional accuracy, precision, recall, and f1-score, affirming the robustness of our feature set and the efficacy of the model in identifying phishing attempts.

Ultimately, our findings contribute to the development of smart tools that can enhance cybersecurity measures and safeguard users from online scams

7. REFERENCES

[1]. APWG. (2023). *Phishing Activity Trends Report*. Anti-Phishing Working Group. https://apwg.org/trendsreports/

[2]. Topkara, M., Kamra, A., Atallah, M., Nita-Rotaru, C., Viwid: Visible Watermarking based Defense Against Phishing. Digital Watermarking, 470–483, September 2005.

[3]. Verma, R., & Das, A. What's in a URL: Fast Feature Extraction and Malicious URL Detection. In *Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics* (pp. 55–63), March 2017

[4]. Patel, D., Shah, V., & Srivastava, S. Machine Learning-Based Phishing Website Detection:A Review and Taxonomy. *Journal of Cybersecurity and Privacy*, *1*(2), 354 374, 2021.

[5]. Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu, LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 3146–3154), December 2017

[6]. Jerome H. Friedman. "Greedy function approximation: A gradient boosting machine.," The Annals of Statistics, Ann. Statist. 29(5), 1189-1232, October 2001

[7]. Friedman, J. H. "Greedy Function Approximation: A Gradient Boosting Machine." Annals of Statistics 29, no. 5, 1189-1232, October 2001