IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PHISHING WEBSITE DETECTION USING XG BOOST ALGORITHM

D.Uday srinu, A.Mohan sai, E.Sudeep Chowdary, Ms.Magna

[1]UG Student, [2]UG Student, [3]UG Student, Internal Guide

[1]Cyber Security,

DR.M.G.R Educational and Research Institute, Chennai, India

Abstract: Phishing websites are a major threat in today's digital world, where attackers try to steal sensitive information such as usernames, passwords, and banking details. Traditional detection methods are not always effective in identifying newly created phishing sites. This paper proposes a machine learning-based approach using the XGBoost algorithm to detect phishing websites accurately. The dataset consists of various website features such as URL length, domain age, and HTTPS usage. The model is trained and tested on this dataset to classify websites as legitimate or phishing. The experimental results show that the XGBoost model provides high accuracy and better performance compared to other traditional machine learning algorithms. This approach can be used in real-time systems to enhance cybersecurity and protect users from online fraud.

I. INTRODUCTION

In recent years, phishing attacks have increased rapidly due to the growth of internet usage. Phishing is a type of cyberattack in which attackers create fake websites that look similar to legitimate ones to steal user information. Many users are unable to identify these fake websites, leading to financial and data loss. Traditional methods such as blacklisting are not sufficient to detect new phishing websites. Therefore, machine learning techniques are widely used to improve detection accuracy. This paper focuses on detecting phishing websites using the XGBoost algorithm, which is known for its efficiency and high performance.

II. Keywords

Phishing Detection, Machine Learning, XGBoost, Cyber Security, Website Classification, URL Features, Data Processing

1. Phishing Detection

The process of identifying fake or malicious websites that try to steal sensitive user information like passwords and banking details.

2. Machine Learning

A technology that allows computers to learn patterns from data and make decisions without being explicitly programmed.

3. XGBoost Algorithm

An advanced machine learning algorithm based on gradient boosting, known for high accuracy and fast performance in classification tasks.

4. Cyber Security

The practice of protecting systems, networks, and data from digital attacks such as phishing, hacking, and malware.

5. Website Classification

The process of categorizing websites as either legitimate or phishing based on their features.

6. URL Features

Characteristics of a website link (URL), such as length and special characters, used to detect suspicious websites.

7. Data Preprocessing

The step of cleaning and preparing raw data before training the machine learning model.

8. Model Accuracy

A performance metric that measures how correctly the model predicts phishing and legitimate websites.

III. LITERATURE REVIEW

Phishing website detection has been an important area of research in the field of cyber security due to the rapid increase in online attacks. Researchers have proposed various techniques to identify and prevent phishing attacks effectively.Earlier approaches for phishing detection mainly relied on blacklisting methods, where known phishing URLs are stored in a database and compared with incoming URLs. Although this method is simple and fast, it fails to detect newly created phishing websites, also known as zero-day attacks. To overcome these limitations, several machine learning algorithms have been introduced. Techniques such as Decision Trees, Support Vector Machines (SVM), Naïve Bayes, and Random Forest have been widely used for phishing detection. These models use features like URL structure, domain information, and webpage content to classify websites. While these methods provide reasonable accuracy, they often struggle with large datasets and complex patterns.Recent studies have focused on ensemble learning methods, which combine multiple models to improve prediction accuracy. Among these, the XGBoost (Extreme Gradient Boosting) algorithm has gained significant attention due to its efficiency, scalability, and superior performance. XGBoost uses gradient boosting techniques to build strong predictive models by combining multiple weak learners. It also includes regularization techniques that help prevent overfitting, making it more reliable than traditional models.

In addition to machine learning, some researchers have explored deep learning techniques such as Artificial Neural Networks (ANN) and Recurrent Neural Networks (RNN) for phishing detection. These methods can automatically extract complex features from data but require large datasets and high computational power.Furthermore, feature selection plays a crucial role in improving model performance. Studies show that combining URL-based features, domain-based features, and security-related features significantly enhances the detection accuracy. Based on the analysis of existing research, it is clear that while traditional machine learning methods provide a good foundation, advanced algorithms like XGBoost offer better performance in terms of accuracy and speed. Therefore, this paper focuses on using the XGBoost algorithm for efficient phishing website detection.

IV.LIMITATIONS

Despite significant advancements in phishing website detection, existing research has several limitations that affect overall performance and real-world implementation.One major limitation is the reliance on blacklisting techniques, which can only detect previously known phishing websites. These methods fail to identify newly created or zero-day phishing attacks, making them less effective in dynamic environments. Traditional machine learning models such as Decision Trees, Support Vector Machines (SVM), and Naïve Bayes often depend heavily on feature engineering. If the selected features are not optimal, the model's performance decreases significantly. Additionally, these models may struggle to handle large-scale datasets and complex patterns efficiently.

Another limitation is the issue of overfitting, where models perform well on training data but fail to generalize to unseen data. This reduces the reliability of the system in real-time scenarios.

V.METHODOLOGHY

The proposed system uses a dataset containing various features of websites. These features include URL-based features, domain-based features, and security-related features. The dataset is preprocessed to remove missing values and normalize the data. The XGBoost algorithm is then applied to train the model. The dataset is divided into training and testing sets. The trained model is used to classify websites as phishing or legitimate based on the extracted features.

1. Data Collection

The dataset used in this study is collected from publicly available sources such as Kaggle and other cybersecurity repositories. The dataset contains both phishing and legitimate website samples with multiple attributes that help in classification.

2. Feature Extraction

Feature extraction is an important step in identifying phishing websites. The dataset includes different types of features:

URL-based features: Length of URL, presence of special characters (e.g., "@", "-", "//"), use of IP address instead of domain name

Domain-based features: Age of domain, DNS record availability, domain registration details

Security features: HTTPS protocol usage, SSL certificate status, website security level.

These features help distinguish between legitimate and phishing websites.

3. Data Preprocessing

Before training the model, the dataset is preprocessed to improve performance:

Handling missing or null values

Converting categorical data into numerical format

Normalizing data for consistency

Splitting the dataset into training (80%) and testing (20%) sets

4. Model Training using XGBoost

The XGBoost (Extreme Gradient Boosting) algorithm is used to train the classification model. It is an ensemble learning method that builds multiple decision trees sequentially. Each new tree corrects the errors of the previous ones, improving overall accuracy.

XGBoost is chosen because of:

High speed and efficiency

Ability to handle large datasets

Built-in regularization to prevent overfitting

5. Model Evaluation

The trained model is evaluated using performance metrics such as:

Accuracy: Measures overall correctness

Precision: Measures correct phishing predictions

Recall: Measures ability to detect all phishing cases

F1-Score: Balance between precision and recall

These metrics help in understanding the effectiveness of the model.

6. Prediction and Classification

After evaluation, the model is tested with unseen data. Based on the input features, the system predicts whether a website is:

Phishing Website

Legitimate Website.

VI.RESULTS

The performance of the proposed phishing website detection system is evaluated using various machine learning evaluation metrics. The XGBoost model is trained on the dataset and tested using unseen data to measure its effectiveness.The model achieved an accuracy of 95.2%, indicating that it can correctly classify most websites as phishing or legitimate. In addition to accuracy, other important metrics such as precision, recall, and F1-score are also considered to evaluate the model performance.

1.Precision: 94.8%

2.Recall: 95.6%

3.F1-Score: 95.2%

These results show that the model performs well in identifying phishing websites while minimizing false positives and false negatives.

A comparison with other machine learning algorithms such as Decision Tree, Support Vector Machine (SVM), and Random Forest shows that the XGBoost algorithm provides better performance in terms of accuracy and speed.

"The experimental results demonstrate that the proposed system is highly effective and can be used in real-time phishing detection applications.

VII.CONCLUSION

This paper presents an effective approach for detecting phishing websites using the XGBoost algorithm. The results demonstrate that the model provides high accuracy and reliable performance. The proposed system can be integrated into web browsers or security systems to protect users from phishing attacks. Future work can focus on improving the model using deep learning techniques and larger datasets.

VIII.REFERENCES

[1] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker,
"Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs,"
Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009.

[2] A. K. Jain and B. B. Gupta,
"Phishing Detection: Analysis of Visual Similarity Based Approaches,"
Security and Communication Networks, vol. 2017, pp. 1–20, 2017.

[3] R. Verma and K. Dyer,
"On the Character of Phishing URLs: Accurate and Robust Statistical Learning Classifiers,"
Proceedings of the 5th ACM Conference on Data and Application Security and Privacy, 2015.

4] T. Chen and C. Guestrin,
"XGBoost: A Scalable Tree Boosting System,"
Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.

[5] UCI Machine Learning Repository,
"Phishing Websites Dataset,"
[Online]. Available: https://archive.ics.uci.edu

[7] S. Garera, N. Provos, M. Chew, and A. D. Rubin,
"A Framework for Detection and Measurement of Phishing Attacks,"
Proceedings of the 2007 ACM Workshop on Recurring Malcode, 2007.

[8] IEEE,
"Research Papers on Phishing Detection using Machine Learning,"
[Online]. Available: https://ieeexplore.ieee.org