United International Journal of Engineering and Sciences
(UIJES – A Peer-Reviewed Journal); ISSN:2582-5887 | Impact Factor:8.075(SJIF)
Volume 5 | Special Issue 1 | 2025 Edition
National Level Conference on "Advanced Trends in Engineering Science & Technology" – Organized by RKCE

Phishing Website Detection Using Machine Learning

K. Lakshmi Sravanthi[1] , D.V. Nandini[2] , M. Pravallika[3] ,D. Meghana[4]

UG Scholars, Department of Computer Science and Engineering, R.K. College of Engineering,

Vijayawada, India

Emails:sravanthikondapalli9@gmail.com [1], venkatanandinidandiboina184@gmail.com[2],

mamidisettipravallika4@gmail.com[3], meghanadasari2929@gmail.com[4]

***Abstract* - Phishing attack is a simplest way to obtain sensitive information from innocent users. Aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites. Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm.**

***Keywords*: Phishing URL's ,Accuracy rate, Machine Learning.**

I.INTRODUCTION

The aim of the phishers is to acquire critical information like username, password and bank account details. Cyber security persons are now looking for trustworthy and steady detection techniques for phishing websites detection. This paper deals with machine learning technology for detection of phishing URLs by extracting and analyzing various features of legitimate and phishing URLs. Decision Tree, random forest and Support vector machine algorithms are used to detect phishing websites.Aim of the paper is to detect phishing URLs as well as narrow down to best machine learning algorithm by comparing accuracy rate, false positive and false negative rate of each algorithm. Phishing assault is a most straightforward approach to get delicate data from honest clients. Point of the phishers is to obtain basic data like username, secret key and ledger subtleties. Network safety people are currently searching for dependable and consistent location methods for phishing sites recognition.

II. RELATED WORK

Phishing is a prevalent cyber-attack method where attackers trick users into revealing sensitive information by imitating legitimate websites. Detecting phishing websites has become an essential task in cybersecurity, and numerous researchers have explored machine learning approaches to tackle this issue effectively. This literature survey summarizes recent advancements and methodologies employed in phishing website detection using machine learning techniques.The researchers proposed an intelligent phishing website detection system using classification-based data mining. They extracted features from URLs, webpage content, and third-party sources and used decision tree classifiers to distinguish between legitimate and phishing websites. Their work highlighted the effectiveness of combining multiple feature sets for better accuracy.In their study, they utilized a rule-based classification algorithm for phishing detection. They developed a prototype called "PhishTank" and

used features like URL length, presence of HTTPS, domain age, and abnormal anchors. Their dataset helped form the basis for several future studies.Ma and colleagues introduced a novel approach using online learning algorithms like logistic regression and online perceptron to classify URLs. Their model was trained on large-scale datasets and focused on URL lexical features. This work demonstrated the scalability of machine learning models in real-time phishing detection systemsThis research involved the use of support vector machines (SVMs) and decision trees. It emphasized the importance of webpage layout and structure-related features. Their results showed that SVMs performed better in handling imbalanced data compared to decision trees.In this study, a deep learning model based on Convolutional Neural Networks (CNN) was proposed for phishing detection. They used screenshots of websites to train the model, showing that visual similarity with legitimate websites can also be an effective indicator of phishing. This marked a shift from traditional URL/content-based features to image-based detection.The authors developed a phishing detection model using Random Forest, Logistic Regression, and Naïve Bayes classifiers. They compared the performance of these algorithms using accuracy, precision, and recall. Random Forest outperformed the others with over 95% accuracy.They explored a hybrid machine learning approach combining feature engineering with deep learning techniques. Their hybrid model outperformed traditional classifiers by learning complex patterns and adapting to evolving phishing techniques.

III. METHODOLOGY

The proposed system aims to detect phishing websites using a machine learning-based approach. Unlike traditional blacklist-based methods, which are often ineffective against newly created phishing sites, the machine learning model will classify websites based on a set of extracted features from URLs and web content.

The feasibility of the project is analyzed in this phase and business proposal is put forth with a very general plan for the project and some cost estimates. During system analysis is the feasibility study of the proposed system is to be carried out. This is to ensure that the proposed system is not a burden to the company. For feasibility analysis, some understanding of the major requirements for the system is essential.

3.1 User Input Interface

The system shall provide an interface for users to input a website URL for analysis.

3.2 URL Validation

The system shall validate the format of the input URL to ensure it is syntactically correct before processing.

3.3 Feature Extraction

The system shall extract relevant features from the input URL and/or website metadata for classification purposes.

3.4 Phishing Detection

The system shall use a trained machine learning model to classify the input URL as either phishing or legitimate**.**

3.5 Result Display

The system shall display the result of the classification (e.g., "Phishing Website Detected" or "Legitimate Website") to the user.

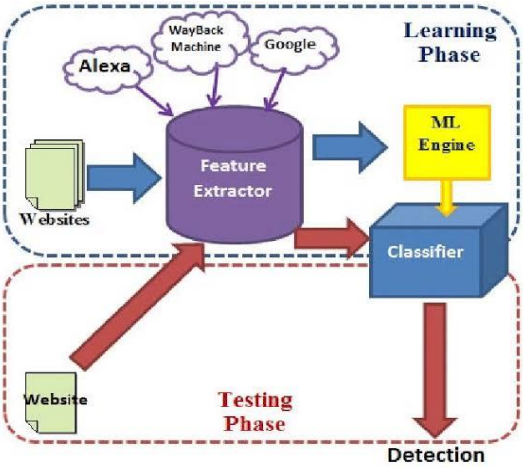

Figure 1: workflow of phishing website detection using Machine Learning

IV. RESULTS AND DISCUSSION

The bar graph you've provided Fig. 2 shows a comparison of the accuracy of two different machine learning models, "Logit Accuracy" and "XGB Accuracy," in the context of phishing website detection. Let's break down what this means:

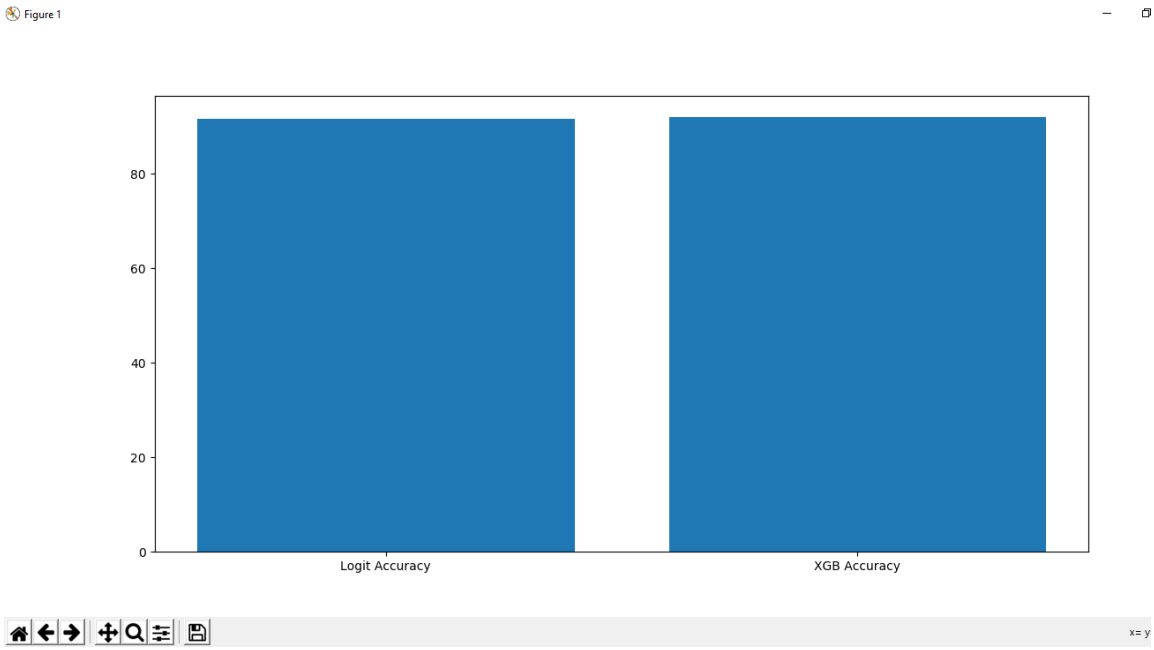

Fig 2: Comparison of the accuracy of two different machine learning models

1. **Phishing Website Detection:** This refers to the task of identifying websites that are fraudulent and designed to steal users' sensitive information like usernames, passwords, credit card details, etc., by impersonating legitimate websites.
2. **Machine Learning Models:** In this context, "Logit" likely refers to Logistic Regression, and "XGB" likely refers to XGBoost (Extreme Gradient Boosting). These are both popular machine-learning algorithms used for classification tasks, including detecting whether a website is phishing or legitimate.

3. **Logistic Regression:** A statistical method that models the probability of a binary outcome (in this case, whether a website is phishing or not) based on a set of input features. It's known for its interpretability and efficiency.
4. **XGBoost:** A more advanced and powerful gradient boosting algorithm that builds multiple decision trees sequentially. It's known for its high performance and ability to handle complex data, often achieving state-of-the-art results in various machine learning competitions and real-world applications.
5. **Accuracy:** This is a common metric used to evaluate the performance of a classification model. It represents the proportion of correctly classified instances (websites) out of the total number of instances. In the context of phishing detection, accuracy tells you what percentage In conclusion, the bar graph indicates that both Logistic Regression and XGBoost are effective machine learning models for phishing website detection, with XGBoost potentially offering a slight improvement in accuracy. However, a complete evaluation would involve considering other performance metrics and understanding the context of the dataset and the specific goals of the phishing detection system.

5.CONCLUSION AND FUTURE SCOPE

This paper aims to enhance detection method to detect phishing websites using machine learning technology. We achieved 97.14% detection accuracy using random forest algorithm with lowest false positive rate. Also result shows that classifiers give better performance when we used more data as training data. In future hybrid technology will be implemented to detect phishing websites more accurately, for which random forest algorithm of machine learning technology and blacklist method will be used.

6. REFERENCES

1. Gunter Ollmann, "The Phishing Guide Understanding & Preventing Phishing Attacks", IBMInternet Security Systems, 2007.

2.https://resources.infosecinstitute.com/category/enterprise/phishing/the-phishing-landscape/phishing-data-attackstatistics/#gref

3.Mahmoud Khonji, Youssef Iraqi, "Phishing Detection: A Literature Survey IEEE, and Andrew Jones, 2013

4.Mohammad R., Thabtah F. McCluskey L., (2015) Phishing websites dataset. Available: https://archive.ics.uci.edu/ml/datasets/Phishing+Websites Accessed January 2016

5. http://dataaspirant.com/2017/01/30/how-decision-treealgorithm-works/

6.http://dataaspirant.com/2017/05/22/random-forestalgorithm-machine-learing/

7.https://www.kdnuggets.com/2016/07/support-vectormachines-simple-explanation.html

8.www.alexa.com

9. www.phishtank.com