iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83793

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Phishing Website Detection using Logistic Regression in Machine Learning

K. Tulasi Krishna Kumar[1], Sheik. Roshini[2]
[1]*Associate Professor & Training & Placement Officer,* [2]*MCA Final Semester, Master of Computer Applications, Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India*

Abstract: ***The rapid growth of internet-based services has significantly increased the risk of cyber threats, among which phishing attacks remain one of the most common and effective social engineering attacks. Phishing websites are designed to deceive users and steal confidential information such as login credentials, banking details, and credit card information. This research focuses on developing an intelligent phishing website detection system capable of accurately identifying malicious and legitimate websites. A comprehensive phishing dataset is collected from sources such as Phish Tank and other verified phishing repositories, containing various website attributes and URL-based features. The collected dataset is pre-processed and analysed using machine learning techniques to classify websites into phishing and legitimate categories. In this study, a Logistic Regression-based machine learning model is employed for phishing website classification due to its efficiency, interpretability, and ability to handle binary classification problems. The proposed model analyses extracted website features and predicts the probability of a website being phishing or legitimate. Additionally, a web-based application is developed to provide real-time phishing detection capabilities for users. The system incorporates a database containing previously identified phishing websites, which functions as a blacklist to improve detection speed and reduce redundant classification operations. Experimental evaluation demonstrates that the proposed Logistic Regression approach provides effective phishing detection performance while maintaining computational efficiency, making it suitable for practical cyber security applications.***

Keywords: Phishing Detection, Machine Learning, Logistic Regression, Multi-Nominal Method, Phish Tank, Website Classification, Blacklist Database.

I. INTRODUCTION

The rapid growth of the internet has increased the number of online services such as banking, shopping, communication, and social networking. Along with these advancements, cyber threats have also become more common, among which phishing attacks are one of the most dangerous. Phishing is a type of cybercrime where attackers create fake websites that imitate legitimate websites to steal sensitive information such as usernames, passwords, credit card details, and personal data from users. Traditional methods of detecting phishing websites mainly rely on blacklists and manual verification. However, these approaches are often ineffective because phishing websites are created and removed quickly, making it difficult to maintain updated databases. To overcome these limitations, Machine Learning (ML) techniques can be used to automatically identify phishing websites by analysing their features and Bhagirathi's project, "Phishing Website Detection Using Machine Learning," aims to develop a system that can accurately classify websites as legitimate or phishing. The system uses various website-related features such as URL length, domain age, HTTPS usage, presence of special characters, redirections, and webpage content characteristics. Machine learning algorithms are trained on datasets containing both phishing and legitimate website information to detect patterns and make predictions. The main objective of this project is to improve cyber security by providing a fast, reliable, and automated method for phishing detection. Different machine learning algorithms such as Decision Tree, Random Forest, Logistic Regression, Support Vector Machine (SVM), and Naive Bayes can be used and compared to determine the most accurate model. This project helps users stay protected from online fraud and contributes to creating a safer digital environment. It also demonstrates how machine learning can be effectively applied in the field of cyber security to solve real-world problems.

II. LITERATURE SURVEY

Many researchers have worked on phishing website detection to protect users from online fraud and cyber-attacks. Earlier, phishing websites were detected using blacklist methods, where website links were checked against a stored database of known phishing sites. However, this method could not identify newly created phishing websites quickly. To overcome this problem, researchers started using Machine Learning techniques.

©IJRASET: All Rights are Reserved

[1] Machine Learning helps computers learn patterns from website data and automatically detect whether a website is safe or fake. Different algorithms such as Decision Tree, Random Forest, Support Vector Machine (SVM), Logistic Regression, and Naive Bayes have been used for phishing detection. These algorithms analysed features like URL length, HTTPS security, domain information, and webpage behaviours to classify websites. Among these methods, Random Forest and SVM often provide better accuracy and reliable results. Recent research also uses Deep Learning techniques to improve detection performance further. From previous studies, it is clear that Machine Learning methods are more effective than traditional methods because they can detect both old and new phishing websites. This project is based on these techniques to create a secure and accurate phishing website detection system.

.

III. CHALLENGES

Developing a phishing website detection system using Machine Learning involves several challenges. One major challenge is detecting newly created phishing websites because attackers continuously change website URLs and designs to avoid detection. Another challenge is collecting accurate and updated datasets, as phishing websites are frequently removed from the internet. Selecting the right features such as URL structure, domain age, and website content is also difficult because some legitimate websites may look similar to phishing websites. [7]Machine Learning models may sometimes produce false results by classifying safe websites as phishing or phishing websites as safe. Handling large amounts of website data and improving the speed and accuracy of the system are additional challenges. Moreover, cybercriminals continuously develop new techniques, so the detection system must be regularly updated to remain effective and reliable.

IV. PROPOSED METHODOLOGY

The proposed system uses Machine Learning to detect whether a website is safe or phishing. First, the system collects website data that includes both real and fake websites. Then, important details such as URL length, HTTPS security, special symbols in the URL, and domain information are taken from the websites. These details are called features. After collecting the features, the data is prepared and divided into two parts: training data and testing data. Machine Learning algorithms like Decision Tree, Random Forest, SVM, and Logistic Regression are trained using the training data. The system learns the differences between safe and phishing websites. When a user enters a website URL, the system checks its features and predicts whether the website legitimate or phishing. This helps users stay safe from online fraud and cyber-attacks. [20]

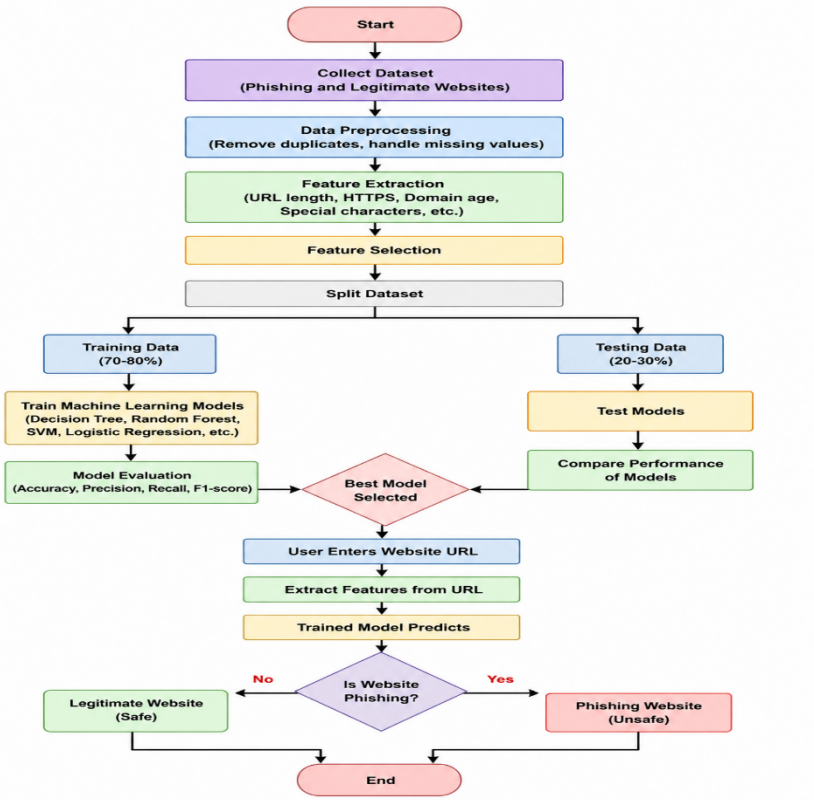


Figure 1: Flow chart of the proposed methodology

V. ALGORITHMS

The phishing website detection system uses different Machine Learning algorithms and techniques to identify whether a website is safe or phishing. [12]These algorithms learn patterns from website data and make predictions based on extracted features.

A. *Decision Tree*

Decision Tree is a supervised machine learning algorithm that makes decisions using a tree-like structure. It checks website features such as URL length, HTTPS usage, and domain details step by step to classify a website as phishing or legitimate. It is simple, fast, and easy to understand.

B. *Random Forest:*

Random Forest is an advanced algorithm that combines multiple decision trees to improve accuracy. Each tree gives its prediction, and the final result is selected based on majority voting. This technique reduces errors and provides better performance for phishing detection.

C. *Support Vector Machine (SVM)*

SVM is a powerful classification algorithm used to separate phishing and legitimate websites by creating a boundary between different data classes. It works effectively with large datasets and gives high accuracy in phishing detection.

D. *Logistic Regression*

Logistic Regression is a statistical machine learning technique used for binary classification, which predicts the probability of a website being phishing or legitimate based on extracted input features. Naive Bayes is another probabilistic machine learning algorithm that applies probability-based calculations to classify websites and provides fast and efficient performance, especially when handling large datasets. The proposed system involves several techniques, including data collection, where phishing and legitimate website URL datasets are gathered from reliable sources for training and testing purposes. Data preprocessing is performed to improve data quality by removing missing values, duplicate records, and irrelevant information. Feature extraction is carried out to identify significant website characteristics such as URL length, HTTPS security status, special characters, and domain age. Feature selection is then applied to choose the most relevant attributes, enhancing the accuracy and efficiency of the classification model. The selected features are used for model training and testing, where machine learning algorithms are trained using training data and evaluated using testing data. Finally, the performance of the proposed system is measured using evaluation metrics such as Accuracy, Precision, Recall, and F1-Score to determine the effectiveness of phishing website detection.

VI. ARCHITECTURE

The phishing website detection system works in two phases: training phase and prediction phase. In the training phase, the system collects data of both phishing and safe websites. The data is cleaned and important features like URL length, HTTPS security, domain age, and special symbols are extracted. Then, Machine Learning algorithms such as Decision Tree, Random Forest, SVM, Logistic Regression, and Naive Bayes are trained using this data. The models are tested, and the best model is selected and saved. [19] In the prediction phase, the user enters a website URL into the system. The system checks the website features and sends them to the trained model. The model predicts whether the website is safe or phishing. Finally, the result is shown to the user to help them avoid unsafe websites and online fraud.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

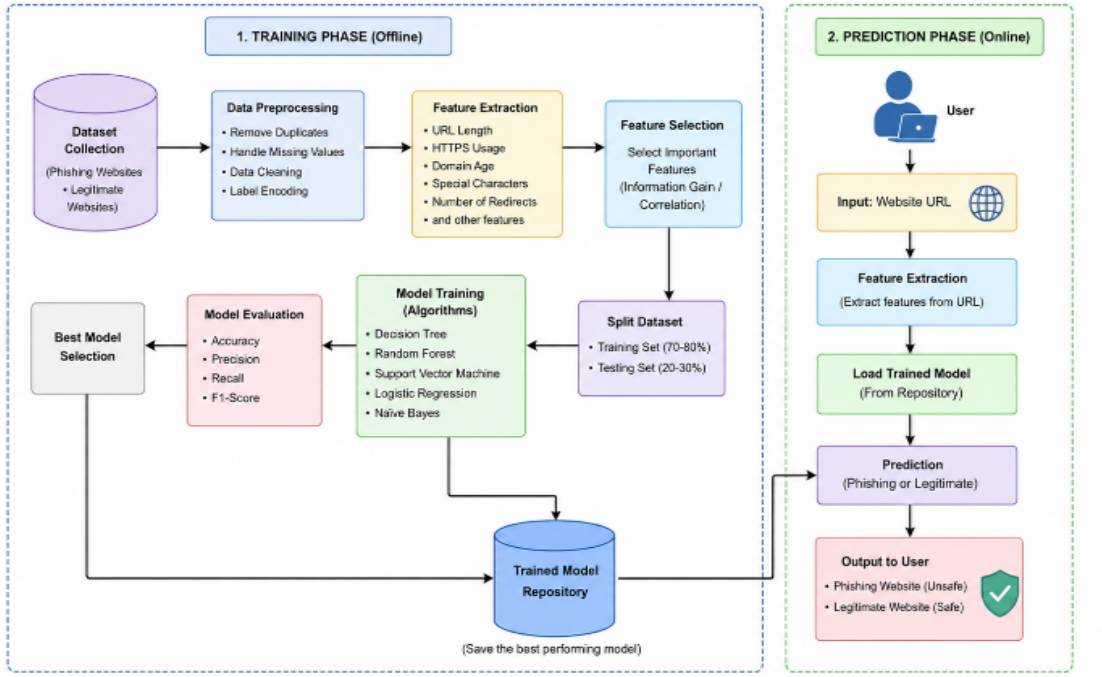


Figure 2: Architecture of the proposed methodology

VII. INPUTS

A. Training phase:

The training phase is the stage where the system learns how to detect phishing websites. First, the system collects data of both phishing and safe websites. Then, the data is cleaned by removing duplicate and missing values. Important website details such as URL length, HTTPS security, domain age, and special characters are extracted. After that, the data is divided into training data and testing data. Machine Learning algorithms like Decision Tree, Random Forest, SVM, Logistic Regression, and Naive Bayes are trained using the training data. The models are tested to check their accuracy and performance. Finally, the best-performing model is selected and saved. [3]This trained model is later used to predict whether a website is safe or phishing in the prediction phase.

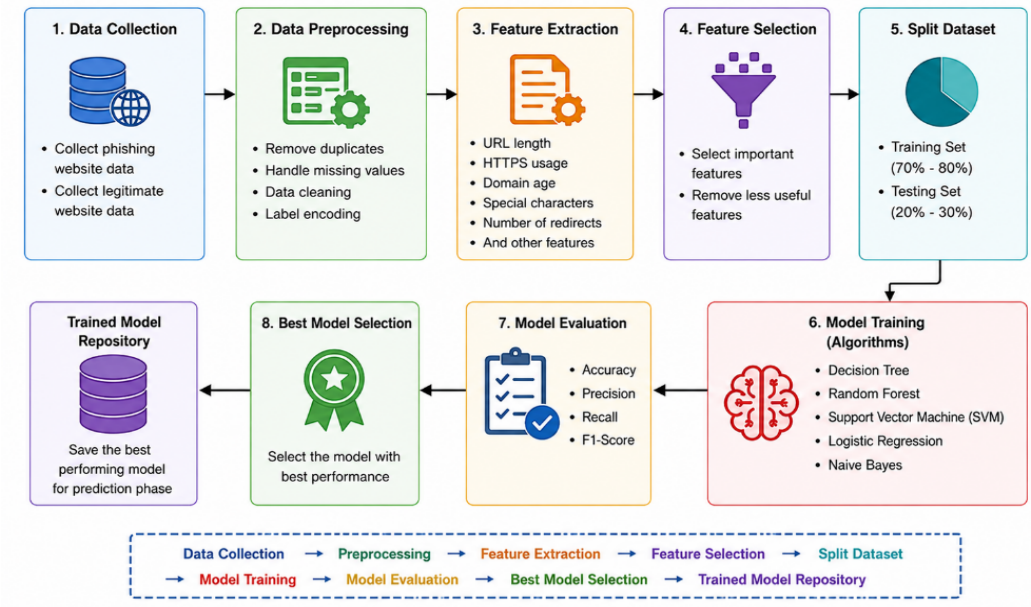


Figure 3: Sample for the Training Phase

B. Prediction phase:

The prediction phase is the stage where the system checks whether a website is safe or phishing using the trained Machine Learning model. First, the user enters a website URL into the system. Then, the system extracts important details from the URL, such as URL length, HTTPS security, and special characters. After extracting the features, the system loads the trained model and analyses the website details. The model compares the information with the patterns it learned during training and predicts whether the website is legitimate or phishing. Finally, the result is shown to the user. If the website is safe, it is displayed as a legitimate website. If it is dangerous, the system warns the user that the website is phishing or unsafe.

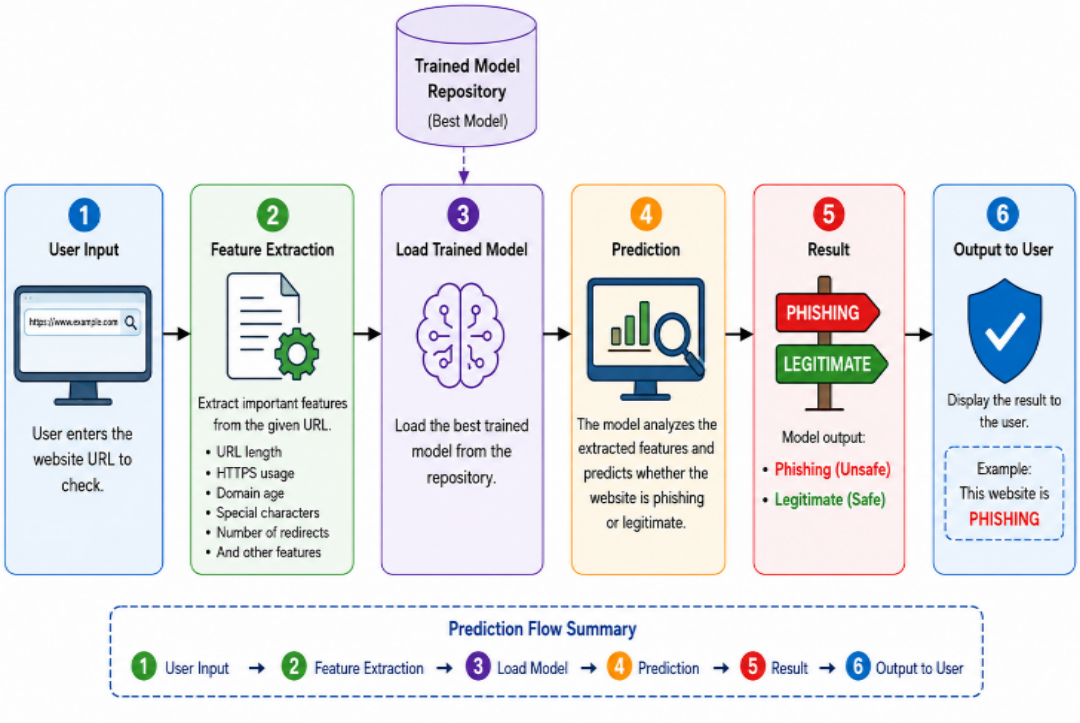


Figure 4: Sample for the Training Phase the Prediction Phase

C. Non-Phishing Website Login Page

A non-phishing website login page is a real and secure webpage used by trusted organizations for user login. These websites use HTTPS security, correct domain names, and secure authentication methods to protect user information. Legitimate websites help users safely access their accounts without the risk of data theft or online fraud users safe access their accounts without the risk of data theft or online fraud.

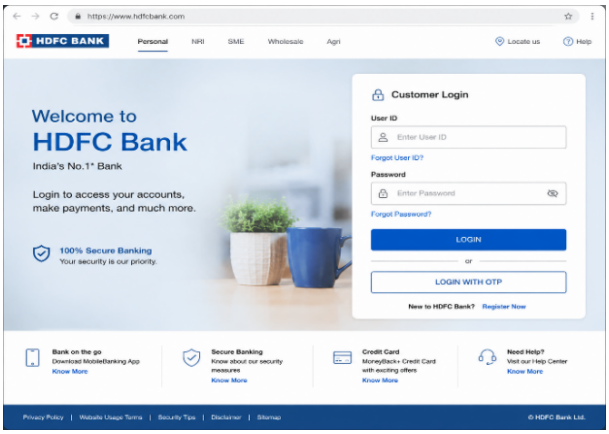


Figure 5: Sample screenshot of Non Phishing Website

D. Phishing Website Login Page

A phishing website login page is a fake webpage created by attackers to steal user information such as usernames, passwords, bank details, and personal data. It is designed to look similar to real websites to trick users into entering their login credentials. These websites often use fake URLs, poor security, suspicious links, and misleading messages. Phishing login pages are dangerous because they can lead to identity theft, financial loss, and cyber-attacks.

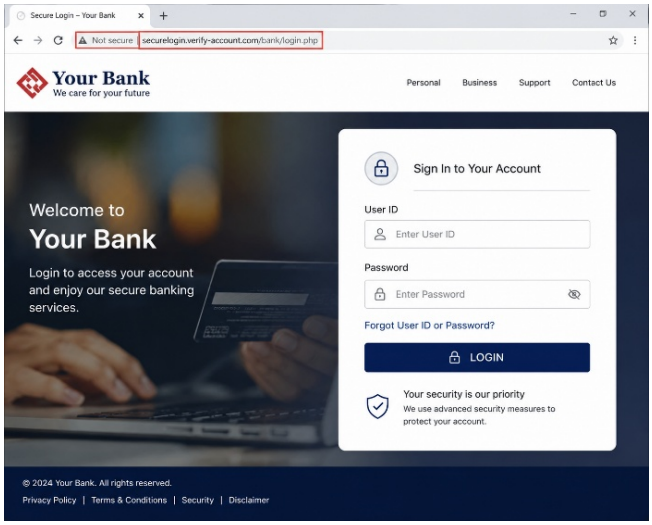


Figure 6: Sample screenshot of Phishing Website

VIII. OUTPUT

A. Output Screen

The output screen displays the final prediction result of the phishing website detection system. After the user enters a website URL, the trained Machine Learning model analyses the website features and shows whether the website is phishing or legitimate. If the website is safe, the system displays a message such as “Safe Website” or “Legitimate Website.” If the website is dangerous, it shows a warning message like “Phishing Website” or “Unsafe Website.” The output screen may also display additional details such as the entered URL, prediction result, confidence score, and checking time. This helps users easily understand the result and avoid accessing fraudulent websites. [9]The output screen improves user awareness and provides better online security.

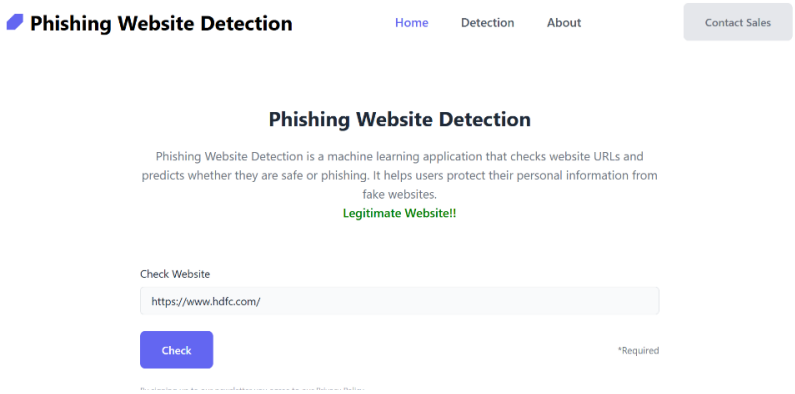


Figure 7: Sample Output screenshot of Non Phishing Website (Reff to Figure 5)

IX. CONCLUSION

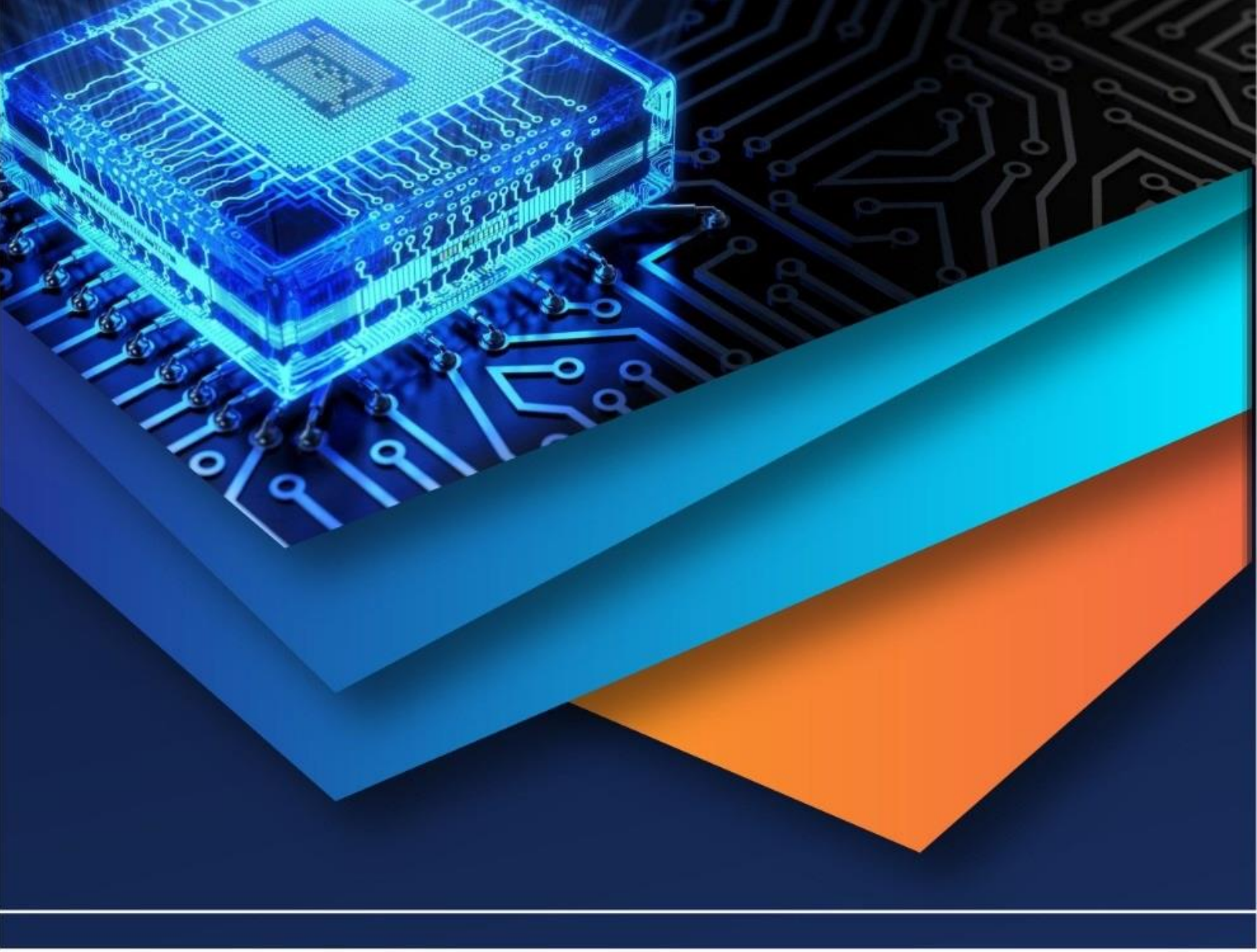
The proposed “Phishing Website Detection Using Machine Learning” system demonstrates an effective approach for identifying and classifying malicious websites by applying machine learning techniques. In this work, the Logistic Regression algorithm is utilized as a binary classification model to analyze extracted website features and predict whether a given website is phishing or legitimate. The system improves the reliability of phishing detection by evaluating important attributes such as URL characteristics, security-related parameters, and domain-based features.

The integration of data preprocessing, feature extraction, and feature selection enhances the performance and efficiency of the classification process. The developed model provides faster detection, reduces the risk of credential theft, and strengthens user protection against cyber threats. Experimental evaluation using performance metrics such as accuracy, precision, recall, and F1-score demonstrates the effectiveness of Logistic Regression in phishing website identification. The proposed system highlights the potential of machine learning-based approaches in developing intelligent cybersecurity solutions and provides a scalable framework for real-time phishing detection and secure web browsing.

REFERENCES

[1] Mohammad, R. M., Thabtah, F., & McCluskey, L. (2015). Phishing Website Detection Using Machine Learning Techniques. International Journal of Cyber-Security and Digital Forensics.
[2] Jain, A. K., & Gupta, B. B. (2018). Towards Detection of Phishing Websites on Client-Side Using Machine Learning Based Approach. Telecommunications Systems Journal.
[3] Abdelhamid, N., Ayesh, A., &Thabtah, F. (2014). Phishing Detection Based Associative Classification Data Mining. Expert Systems with Applications.
[4] Chiew, K. L., Yong, K. S., & Tan, C. L. (2018). A Survey of Phishing Attacks: Their Types, Vectors and Technical Approaches. Expert Systems with Applications.
[5] Rao, R. S., & Pais, A. R. (2019). Detection of Phishing Websites Using an Efficient Feature-Based Machine Learning Framework. Neural Computing and Applications.
[6] Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019). Machine Learning Based Phishing Detection from URLs. Expert Systems with Applications.
[7] Verma, R., & Das, A. (2017). What's in a URL: Fast Feature Extraction and Malicious URL Detection. ACM International Conference Proceedings.
[8] Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs. ACM SIGKDD Conference.
[9] Xiang, G., Hong, J., Rose, C. P., & Cranor, L. (2011). CANTINA+: A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites. ACM Transactions on Information and System Security.
[10] Basnet, R., Mukkamala, S., & Sung, A. H. (2012). Detection of Phishing Attacks: A Machine Learning Approach. Soft Computing Applications.
[11] Scikit-learn Documentation –Machine Learning Library for Python. Web link: Scikit-learn Official Website
[12] Python Programming Language Documentation.Web link: Python Official Website
[13] TensorFlow Documentation – Deep Learning Framework. Web link: TensorFlow Official Website
[14] Kaggle – Phishing Website Dataset. Web link: Kaggle Official Website
[15] UCI Machine Learning Repository – Phishing Websites Dataset.UCI Machine Learning Repository
[16] WEKA Data Mining Tool Documentation. Web link: WEKA Official Website
[17] National Institute of Standards and Technology (NIST) Cybersecurity Resources. Web link: NIST Official Website
[18] Open Phish – Phishing Intelligence and Feed Services. Web link: Open Phish Official Website
[19] Phish Tank – Community Phishing Detection Platform. Web link: Phis Tank Official Website
[20] Google Safe Browsing – Website Security Service. Web link: Google Safe Browsing

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)