iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82896

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 V May 2026- Available at www.ijraset.com

Real-Time Phishing Website Detection Using Machine Learning Techniques

Prof. S.L.Tambe[1], Om Sudhakar Chaudhari[2], Khushal Sunil Lohar[3], Kundan Gajanan Patil[4], Prathamesh Sanjay Khairnar[5]

[1, 2, 3, 4]*UG Students, R.C. Patel Institute of Technology, DBATU, Shirpur, India*

[5]*Project Guide, R. C. Patel Institute of Technology, DBATU, Shirpur, India*

Abstract: ***At present, phishing websites pose one of the greatest cybersecurity threats on the Internet. These websites are fraudulent sites that try to fool users into sharing passwords, bank details, and other sensitive data by imitating legitimate websites. Since cybercriminals keep coming up with phishing websites, the task of identifying them through traditional means becomes progressively challenging. This is why machine learning can serve as a novel approach for phishing detection. In this research paper, we propose a machine learning model to detect phishing sites based on various attributes associated with the websites and URLs, like URL length, domain age, use of HTTPS, redirections, and presence of suspicious characters in website URLs. Some of the popular machine learning techniques, namely, Random Forest, SVM, Decision Tree, Logistic Regression, and ANN, are used to evaluate their effectiveness in detecting phishing. The experiments have been performed by applying the algorithm to public datasets from reliable sources like Phish Tank and the UCI Machine Learning Repository. It can be observed from the results obtained that machine learning algorithms are capable of effectively identifying phishing websites from legitimate websites with enhanced accuracy. This system has the potential to aid in enhancing cybersecurity strategies and detecting phishing websites in real time.***

Keywords: Phishing Websites Identification, Machine Learning, Cybersecurity, URLs Analysis, Random Forest, Support Vector Machines (SVM), Artificial Neural Networks (ANN), Websites Classification, Phishing Attacks, Real-Time Detection.

I. INTRODUCTION

The internet has grown to be an important aspect in our lives due to a variety of ways including communication, banking, buying things, studying, and others. With increasing uses of the internet, there is also an increase in threats associated with cyberspace at a substantial level. The most common type of cyber attack is called phishing. In this kind of attack, the criminals make web pages that are similar to those of genuine sites to fool people into divulging their credentials.

Detection of phishing scams through the usual means depends greatly on blacklist database systems and rule-based security mechanisms, which have been manually created. They might detect any old phishing sites, but new ones could go undetected since the usual method would not apply as the attackers find new ways to launch attacks on the security system.

The use of machine learning techniques for detecting phishing websites is one of the most efficient methods due to its ability to recognize hidden patterns within data and detect suspicious activities automatically. Machine learning models are capable of recognizing URL-based characteristics, domain attributes, web page content, and other security factors to determine whether a website is a phishing website or not.

In this research paper, the design and evaluation of a machine learning-based phishing websites detection system is carried out. Several machine learning models are evaluated to determine the best methodology that can be employed to detect phishing attacks effectively.

Due to increased usage of the internet in our modern society, the existence of phishing websites constitutes a major cyber-security threat nowadays. Phishing websites usually resemble normal websites and aim to fool the user into divulging personal information including user names, passwords, banking credentials, as well as personal information. Conventional phishing websites detection methods, which include blacklisting and rules-based approaches, cannot efficiently recognize new phishing websites.

Machine learning technology constitutes a smart tool that has been developed and applied in order to detect phishing websites more intelligently. Machine learning algorithms make use of different features including URL length, HTTPS usage, domain age, URL redirects, and suspicious characters among other things, in determining whether or not the website under consideration is a phishing website.

II. OBJECTIVES

1) For comprehending the workings and implications of phishing sites in today's cybersecurity frameworks.
2) For developing an efficient intelligent system using machine learning for detecting phishing sites effectively.
3) For collecting and preprocessing datasets of phishing sites and legitimate sites to train and test models.
4) For analyzing significant features of URLs and websites, including URL length, HTTPS, redirects, and domain age..
5) For implementing various machine learning techniques, such as Random Forest, SVM, Decision Tree, Logistic Regression, and ANN.
6) For enhancing phishing site detection efficiency while minimizing false positives and negatives.

III. LITERATURE REVIEW

Several studies have highlighted the use of various forms of machine learning and deep learning for the detection of phishing sites. The earlier techniques of phishing detection involved the use of blacklisting and rule-based approaches, which were unable to identify new phishing sites. In order to address this issue, many researchers have used machine learning methods for detecting phishing sites.

Safi, A., and Singh, S. conducted an analysis on the application of machine learning in phishing web site detection and found that machine learning provided better accuracy compared to blacklisting approaches. According to Safi and Singh, algorithms including Random Forest, SVM, and Decision Trees exhibited significant performance in detecting phishing web sites using URL and Webpage features [1].

According to Rao & Pais, the application of machine learning in detecting phishing web sites was first introduced by algorithms such as Random Forest, Logistic Regression, and Decision Trees. However, they concluded in their research that the algorithm of Random Forest gave better results because it had the ability to deal with large datasets [2].

Choudhary and Jain conducted their research on the design of a machine learning model for phishing attack detection using the relevant attributes from the URLs, such as the age of the domain, length of the URL, presence of redirections, and suspicious characters. The model succeeded in classifying phishing websites correctly [3].

Another research done by Ali demonstrated the application of a supervised machine learning technique with feature selection via wrappers for phishing site detection. Based on their results, the choice of essential features for the website increased the accuracy and effectiveness of machine learning approaches [4].

In some recent studies, there has been an emphasis on using deep learning models and hybrid approaches for the phishing attacks. In one of the researches conducted by Gupta et al., the authors introduced a hybrid approach for phishing detection through feature selection. As per their results, hybrid models outperformed single machine learning models [5].

Approach for phishing URL detection was suggested by Rehman et al. using machine learning models. The importance of employing these detection mechanisms in cybersecurity tools in today's technology was highlighted. It was revealed through the study that ensemble learning models perform very effectively in phishing site detection.[6]

Therefore, through research so far carried out on the topic, it has become clear that the application of machine learning algorithms as well as hybrid methods in phishing website detection are highly effective. But there are some areas which require further improvements; these include real-time phishing detection, zero-day phishing detection, and reduction of false positive rates.

IV. METHODOLOGY

A phishing website detection approach that will be adopted is designed based on machine learning techniques. The methodology involved in building this system includes various essential steps such as the collection of data, data preprocessing, feature extraction, training the model, making predictions, and evaluating the results. The entire project was developed by using Python, Flask, Scikit-learn, and SQLite database technologies.

A. Data Collection

Phishing websites detection using the phishing website dataset will be employed in this project. This dataset will serve as both the training and testing set for the machine learning model. It is available in CSV format and comprises 11,055 instances of website data with 30 predictor attributes and one response variable known as "Result."

The data set includes various URL and Website features like URL Length, HTTPS usage, redirects, age of domain, unusual symbols, DNS entries, traffic on the website, number of sub-domains, and details about Google indexing. These features assist in helping the machine learning model recognize phishing websites better.

The target values used in the dataset are represented as:
1 = Legitimate Website
0 = Suspicious Website
-1 = Phishing Website
The dataset is processed using the Pandas library, and preprocessing techniques such as feature separation and train-test splitting are applied before model training. The dataset is divided into training and testing sets using an 80:20 ratio to evaluate the performance of the machine learning model effectively.

B. Feature Engineering

One of the key processes in the developed phishing website detection system is feature extraction. Here, various attributes related to URLs and websites are examined and encoded into numerical values that can be comprehended by the machine learning algorithm. Feature extraction helps determine the authenticity of the websites based on their behavior and structures as legitimate, suspicious, or phishing sites.
The designed system employs 30 phishing-specific features for each URL submitted to the website. These features comprise URL-based, domain-based, and security-related attributes that have traditionally been used in phishing website detection systems.
The following are some of the critical extracted features:

- IP Address: Checks whether the website uses an IP address instead of a domain name, which is commonly done by phishing sites.
- URL Length: Detects whether the length of the URL is unusual or suspicious.
- URL Shortening Service: Detects the use of URL shortening services to disguise malicious links.
- Presence of '@' Symbol: Detects unusual symbols used in phishing URLs.
- Double Slash Redirecting: Detects the presence of unnecessary redirects in the URL.
- Abnormal Prefix or Suffix: Detects the use of abnormal hyphens in domain names.
- Subdomains Used: Determines how many subdomains are used in the URL.
- HTTPS/SSL State: Checks whether the website uses secure HTTP protocols.
- Duration of Domain Registration: Checks how long a domain registration lasts; phishing sites generally have very short registrations.
- Request URL and URL of Anchor: Detects suspicious links that try to access external resources.
- DNS Record Existence: Determines whether the site has valid DNS records.
- Web Traffic & Google Index: Detects whether the site gets any traffic or whether Google indexes it.
- Age of Domain: Determines how long the domain exists; phishing sites generally use recently registered domains.
- Statistical Indicators: Detects phishing-related statistical indicators.

The extracted features are formatted through Pandas Data Frame before being used in the machine learning model for predicting and classifying purposes.
The process of feature extraction enhances the capacity of the system to identify phishing websites through considering various security features rather than basing the identification process on blacklisting techniques.

C. Data Preprocessing

Preprocessing is a crucial process in the proposed Phishing Website Detection System because it assists in preparing the data to be used in developing the predictive model. Preprocessing enhances data quality by eliminating inconsistencies and transforming the data into a format that can be classified using algorithms. Preprocessing in this project involves the use of Python packages like Pandas and Scikit-learn.
At first, the dataset containing the phishing websites is loaded into the program from the CSV file with the help of the Pandas library. The dataset consists of 11,055 entries with 30 attributes and one result column called "Result". Prior to building the predictive models, the system performs a check for null values in the dataset.
Next, the input features and output values are segregated. The target column 'Result' consists of classifications, and the other columns consist of the input features. The output/target values will be given as follows:

1 – Legitimate Website
0 – Suspicious Website
-1 – Phishing Website

Following the segregation of the input features and target values, splitting the dataset occurs via the train-test split method. In this particular project, the dataset is split into an 80-20 ratio, where 80% of the data will serve as training data and the remaining 20% as testing data.

Preprocessing makes sure that the data set is clean and structured before features can be extracted from it and machine learning models trained on it. This increases the probability of accurate classifications being made.

D. Model Development

Machine Learning Models

- Random Forest Classifier: The method uses several decision trees to enhance phishing website classification accuracy.
- Support Vector Machine (SVM): It employs the Radial Basis Function (RBF) kernel for separating and finding complex patterns among phishing and non-phishing websites.
- Decision Tree: This classifier classifies websites based on their various features related to phishing attacks.
- Logistic Regression: The model classifies a website as phishing or not using probability techniques.
- Artificial Neural Network (ANN): This algorithm has multiple neurons to detect complex phishing websites.
- Hybrid Approach: A combination of several machine learning methods used in classifying phishing websites more accurately and reliably.

E. Forms Saving and integration

The trained models are all saved and loaded request:

- ML models saved in .pkl format

F. Web Application Development

It is worked out using flask and where:

- The proposed phishing website detection system is developed as a web-based application implemented in Flask framework of Python programming language.
- The front end of the web application is developed using HTML5, CSS3, and Jinja2 templates for simplicity.
- The back end processes URL, performs feature extraction, makes predictions using the machine learning algorithm, performs database operation, and prepares output results using the Flask web application framework.
- User input URL in the application will be analyzed as Legitimate, Phishing, and Suspicious type websites.
- The feature extraction process includes various aspects related to the phishing website and performs analysis of different features associated with a URL before making any prediction by the machine learning algorithm.
- The machine learning model makes predictions based on the trained Random Forest Classifier, giving a prediction result with a confidence score.
- The web application uses the SQLite database system to store scanning information along with blacklisted suspicious URL records.
- The Admin Panel is built-in that manages scanning history, manages the blacklist list of suspicious URLs, and tracks phishing detection statistics.

G. Prediction real time workflow

- A user opens the phishing website detection web application from a web browser.
- The user then inputs a URL in the input box that is present on the homepage of the web application.
- The Flask backend takes the inputted URL and first validates the input.
- The system then identifies whether the inputted URL already exists in the blacklist database. If the URL exists in the database, it will be classified as a phishing website instantly
- However, in case the inputted URL is not listed in the database, the feature extraction module extracts the features of the phishing website URL which include URL length, use of HTTPS, redirects, domain age, suspicious characters, DNS records, and subdomains.
- The extracted features are converted into a Pandas Data Frame for machine learning prediction.
- The random forest algorithm predicts whether the website is Legitimate, Suspicious, or Phishing based on the extracted features.

- Additionally, the prediction is carried out using probability-based classification methods where the prediction is associated with confidence score.
- The output from the machine learning prediction process and confidence score is displayed to the user in real-time on the result page.
- Finally, the data is stored in the SQLite database for future reference and analysis.

H. Model Evaluation

The tools used to evaluate the trained models include:

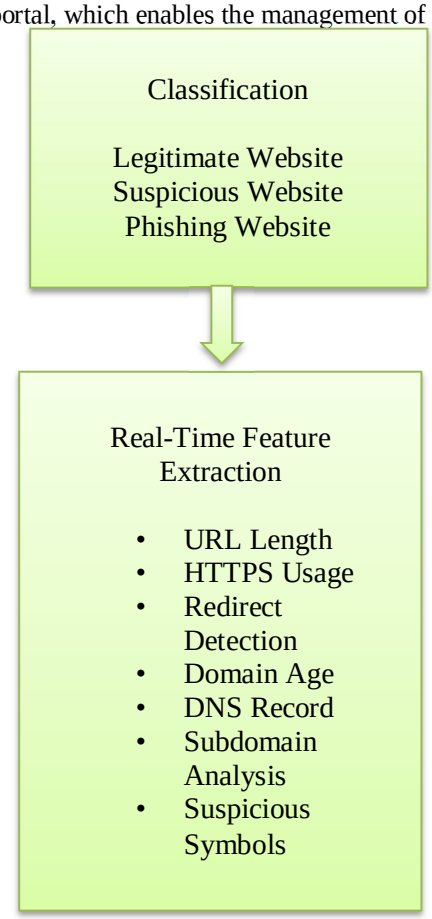
- Accuracy
- Confusion Matrix
- Precision, Recall, F1 Score

I. Summary

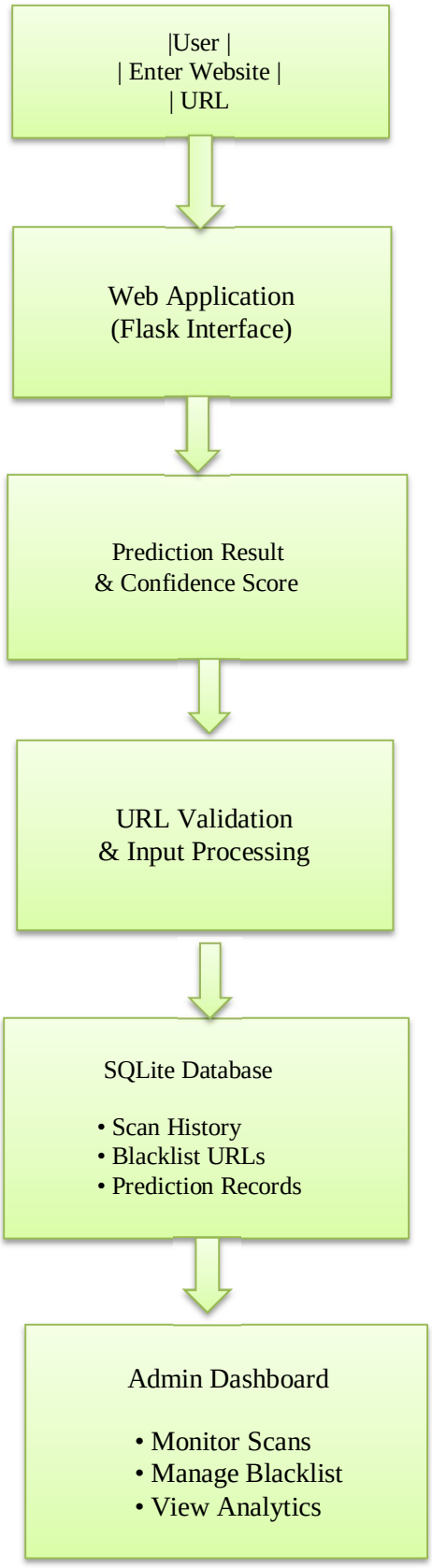
Phishing websites have emerged as one of the significant challenges in terms of cybersecurity in the current digital environment. Such fraudulent websites appear like genuine websites and are primarily used to capture sensitive user data such as login passwords, banking information, and other personal data. However, the traditional methods of phishing detection do not work effectively due to the constant creation of phishing websites that are not easy to detect through a blacklist and rule-based method. In this regard, the use of machine learning for phishing website detection was proposed to tackle the problem.

The system is developed using Python, Flask, Scikit-learn, Pandas, and SQLite technologies. The web application analyzes significant parameters associated with websites and URLs, including the length of URL, use of HTTPS, redirection, suspicious characters, and domain age, to detect any phishing attack. Furthermore, the Random Forest algorithm has been chosen for classification purposes as it offers high precision and reliability.

In this project, a user can input any website URL to obtain prediction results, which include Legitimate, Suspicious, or Phishing. Also, the system consists of an administrator portal, which enables the management of the scan history and blacklisted websites..



- Block Diagram

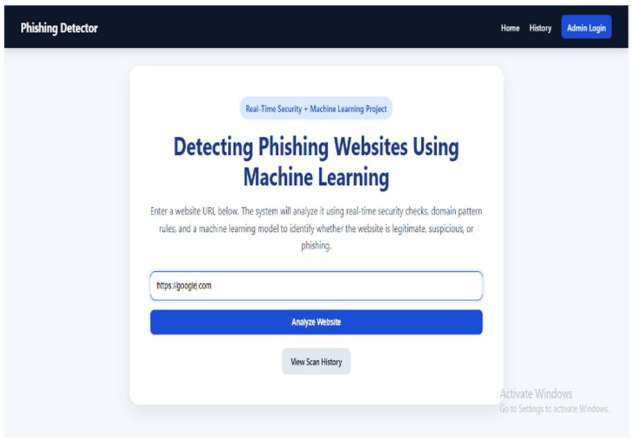


V. RESULT

The phishing website detection system that was proposed was successfully deployed as a web application in real time utilizing machine learning. The web application examines URLs and categorizes them as either Legitimate, Suspicious, or Phishing based on various phishing-based features like the length of the URL, HTTPS utilization, redirects, domain age, and suspicious characters in the URL. The performance of the model, which was trained using the Random Forest Classifier, was excellent, and the results were accurate and dependable. The application provides real-time scores for predictions and saves history scans into an SQLite database..

Step 1: Accessing the Web Application

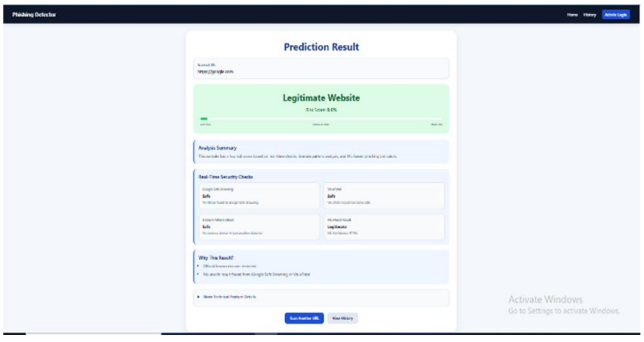
The first step involves accessing the system in a web browser through the project URL. Upon entering the URL into the browser, the landing page of the web application appears, containing a field where users can enter the URL of a website.



Step 2: Input Website URL

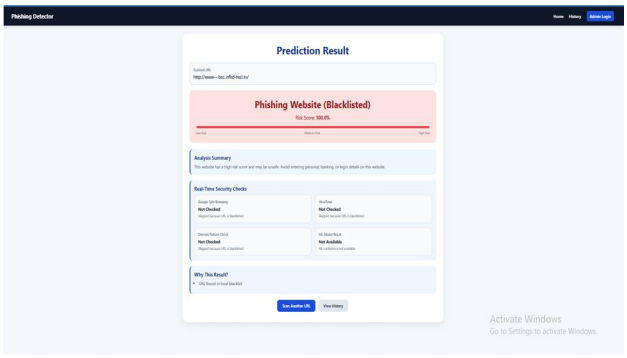
The user inputs the website URL within the designated box. The tool can detect phishing attacks on any website URL.

Example: https://google.com



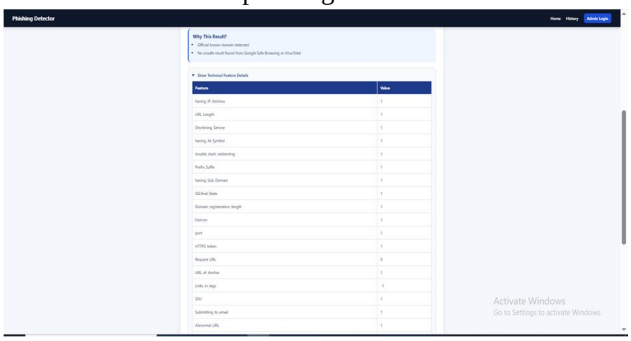
Step 3 :- Select “Analyze Website”

Once the URL has been inputted, the user selects the “Analyze Website” option. This will send the URL inputted to the Flask backend server for analysis.



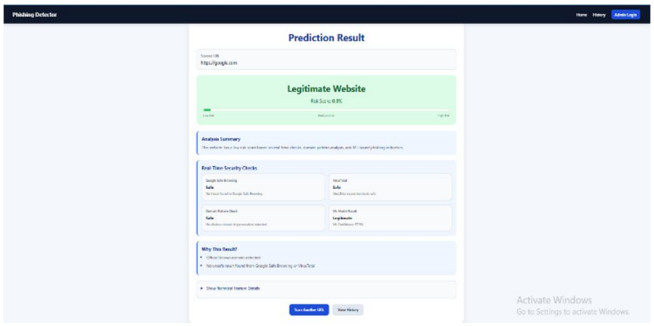
Step 4: Validate URL and Check Against Blacklist Database

The system validates the URL inputted by the user and determines if the site has been flagged as part of the blacklist database. If the URL is blacklisted, the system categorizes the website as a phishing website.



Step 5: Real-Time Feature Extraction

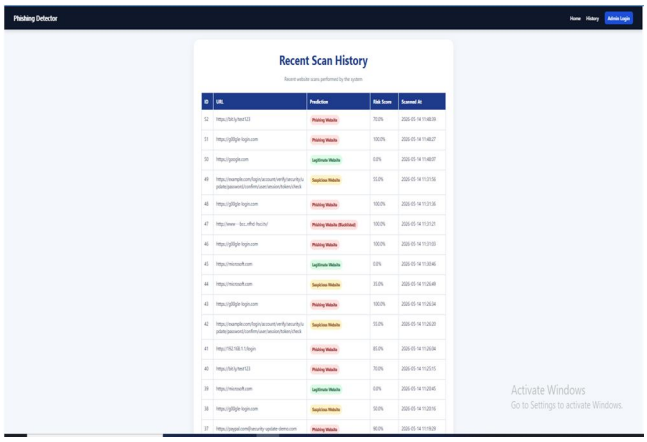
If the URL is not on the blacklist, the feature extraction stage extracts various features related to phishing such as:
Length of the URL
Whether there is an HTTPs present
Redirects detection
Domain age
DNS information
Suspicious characters
Sub-domains data



Step 6: Predictive Machine Learning Algorithm

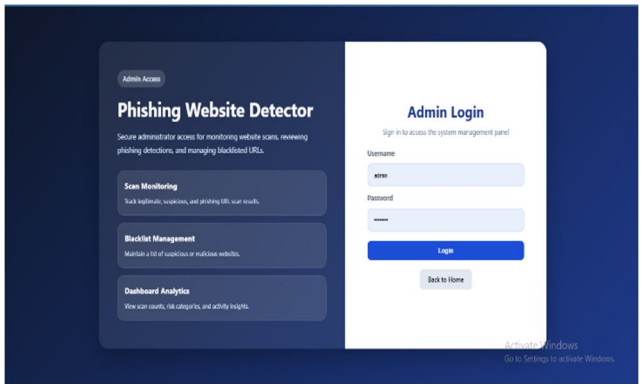
The features are then used for analysis using the predictive Random Forest machine learning algorithm which classifies the website as:

- Legitimate Website
- Suspect Website
- Phishing Website



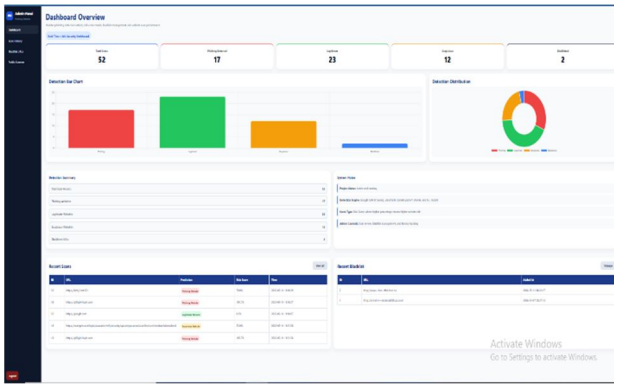
Step 7 : Database Storage of Scanned History

Everything including the scanned URL, prediction result, and the level of accuracy would be stored in the database through the use of SQLite.



Step 8 : Administrator Login

The administrator logs into the system by logging into the admin login page, after which the username and password are entered.



Step 9 : Admin Dashboard Monitoring

On logging in successfully, the admin dashboard shows the following information:

- Scan history
- Phishing stats
- Blacklisted URLs
- Detection analysis
- Search & filter

The administrator can track phishing activity and manage blacklists through the dashboard.

VI. CONCLUSION

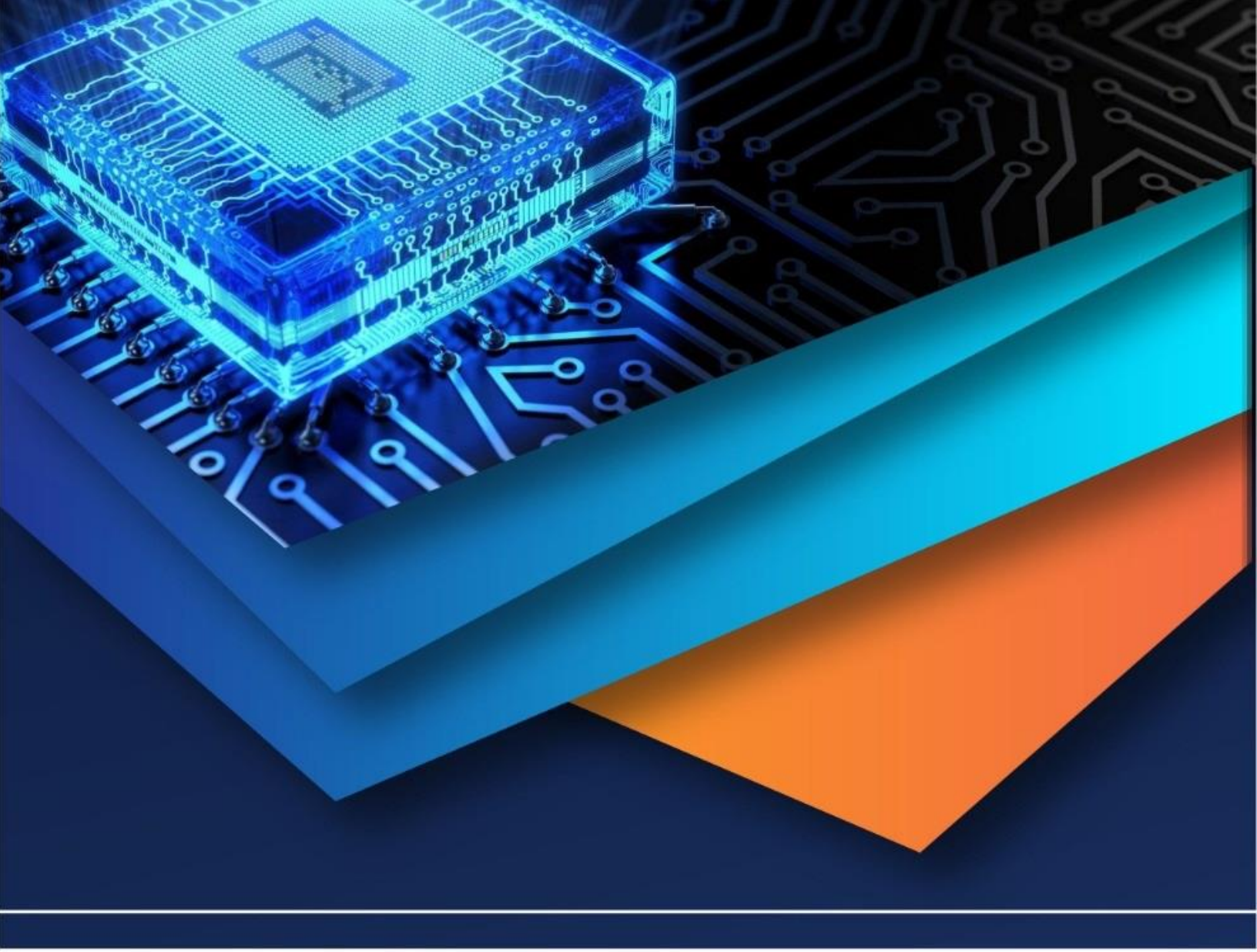
The phishing website detection system that has been suggested is a very efficient solution for detecting phishing websites through machine learning methods. The system efficiently identifies different characteristics related to the URL and website such as URL length, HTTPS, redirectors, domain age, and other suspicious characters to categorize a website as a legitimate one, a suspicious one, or a phishing site. The Random Forest Classifier provided efficient results and performed well in the prediction phase of the project. In addition to these, the web application allows real-time predictions, scanning history records, blacklist functions, and admin monitoring features

REFERENCES

[1] Safi, A., & Singh, S., "A Systematic Literature Review on Phishing Website Detection using Machine Learning Techniques," Journal of King Saud University - Computer and Information Sciences, 2023.
[2] Rao, M. A., & Pais, B., "Phishing Website Detection using Machine Learning Algorithms," International Journal of Computer Applications, 2019.
[3] Choudhary, T., & Jain, S., "A Machine Learning Approach for Phishing Attack Detection," Journal of Artificial Intelligence and Technology, 2023.
[4] Ali, W., "Phishing Website Detection based on Supervised Machine Learning with Wrapper Features Selection," International Journal of Advanced Computer Science and Applications, vol. 8, no. 9, 2017.
[5] Gupta, S. D., et al., "Modeling Hybrid Feature-Based Phishing Websites Detection Using Machine Learning Techniques," Computers & Security, 2022.
[6] Rehman, A. U., et al., "Real-Time Phishing URL Detection Using Machine Learning," Engineering Proceedings, vol. 107, no. 1, 2025.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY