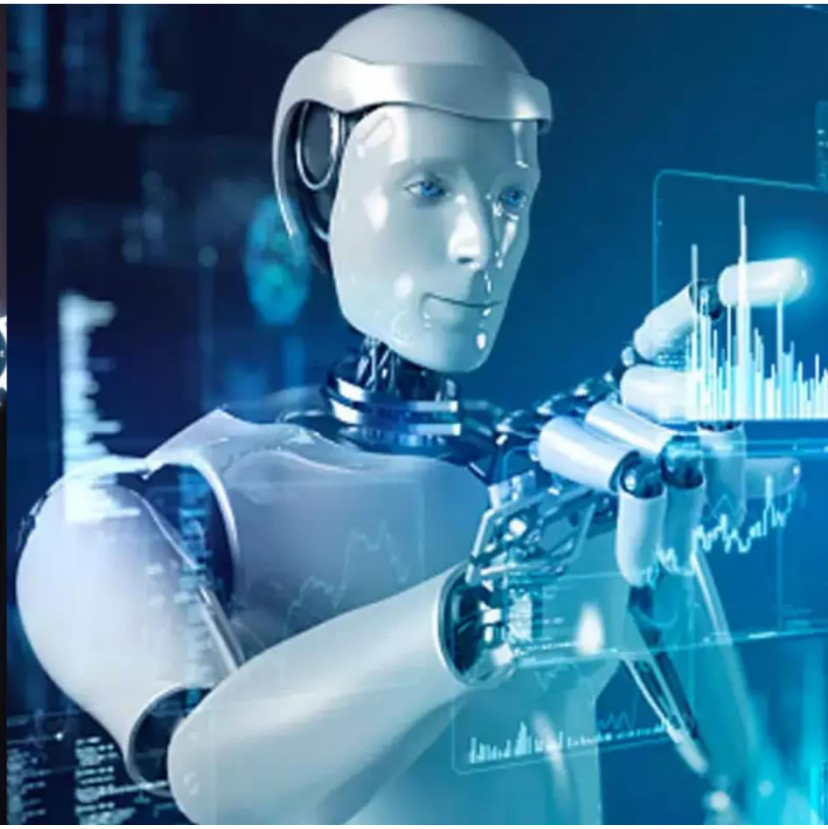
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 3, March 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503394|

Phishing Website Detection using Machine Learning

Prof. K. Jaya Bharathi, G. Revathi, P. Ashwitha, Ch. Sriram

Professor, Department of IT (Information Technology), ACE Engineering College, Hyderabad, India

IV B. Tech Students, Department of IT (Information Technology), ACE Engineering College, Hyderabad, India

ABSTRACT: The rapid growth of internet usage has significantly increased cyber threats, especially phishing attacks. Phishing websites are designed to steal sensitive information such as usernames, passwords, and personal data from users. Traditional detection methods often fail to identify new and evolving phishing techniques and may produce inaccurate results. This project proposes a Phishing Website Detection System using Machine Learning to improve detection accuracy. The system analyzes various URL-based features such as URL length, presence of HTTPS, and special characters. A trained machine learning model is used to classify whether a website is safe or phishing. In addition, validation techniques are integrated to enhance system reliability. These include URL format validation, domain existence checking, and website availability verification. The system is implemented using Flask to provide an interactive and user-friendly web interface. Users can input a URL and receive real-time results with safety percentage and status messages. The combination of validation techniques and machine learning helps reduce incorrect predictions. Overall, this system provides an effective solution for detecting phishing websites and improving online security

I. INTRODUCTION

The rapid growth of internet usage and digital services has led to a significant increase in cyber threats, particularly phishing attacks, which aim to steal sensitive information such as login credentials, financial data, and personal details. As users increasingly rely on online platforms for communication, banking, and transactions, ensuring website security has become a critical concern. Traditional rule-based and basic machine learning approaches for phishing detection often lack accuracy and fail to detect newly evolving phishing techniques. Moreover, these systems may produce incorrect predictions when handling invalid or non-existent URLs, reducing user trust and system reliability. To address these challenges, this project proposes a Machine Learning-based Phishing Website Detection System integrated with real-time URL validation techniques. The system analyzes various URL-based features such as length, presence of HTTPS, domain characteristics, and special symbols to identify potential phishing patterns. In addition to machine learning prediction, the system incorporates validation mechanisms including URL format verification, domain existence checking, and website availability testing to ensure reliable results. The model is deployed using a Flask-based web application, allowing users to input URLs and receive instant feedback with safety percentage and status. By combining intelligent prediction with validation techniques, the system improves detection accuracy, reduces false results, and enhances user confidence. This project demonstrates an effective and practical approach to identifying phishing websites and strengthening cybersecurity in modern web environments.

II. EXISTING SYSTEM

The existing phishing website detection systems primarily rely on traditional rule-based methods and basic machine learning techniques to identify malicious URLs and protect users from cyber threats. These systems analyze features such as blacklisted URLs, keyword matching, domain age, and simple URL characteristics to detect phishing websites. Commonly used approaches include blacklist-based detection, heuristic rules, and machine learning models such as Logistic Regression, Decision Trees, and Random Forests. While these methods can identify known phishing patterns and improve detection accuracy to some extent, they still have several limitations. Many systems fail to detect newly created or sophisticated phishing websites that do not exist in blacklists, making them less effective against evolving attacks. Additionally, traditional models often produce incorrect predictions when handling incomplete, invalid, or non-existent URLs, reducing system reliability. Some systems also lack proper validation mechanisms, leading to misleading outputs such as marking broken or fake URLs as safe. Furthermore, these approaches may not provide clear

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503394|

differentiation between invalid inputs and actual phishing websites, which affects user understanding. As a result, existing systems are not fully reliable, lack robustness, and require improvement to handle real-time scenarios more effectively.

III. PROPOSED SYSTEM

The existing phishing website detection systems primarily rely on traditional rule-based approaches and basic machine learning techniques to identify malicious URLs and protect users from cyber threats. These systems analyze various features such as blacklisted URLs, keyword matching, domain age, and simple URL characteristics to detect phishing websites. Commonly used methods include blacklist-based detection, heuristic rules, and machine learning models such as Logistic Regression, Decision Trees, and Random Forests. While these approaches are effective in identifying known phishing patterns and improving detection accuracy to a certain extent, they suffer from several limitations. Many systems fail to detect newly created or advanced phishing websites that are not present in blacklists, making them less effective against evolving attacks. Additionally, traditional models often produce incorrect predictions when dealing with incomplete, invalid, or non-existent URLs, which reduces system reliability. In many cases, the absence of proper validation mechanisms leads to misleading results, such as classifying broken or fake URLs as safe. Furthermore, these systems do not clearly distinguish between invalid inputs and actual phishing websites, which can confuse users. As a result, existing systems lack robustness, accuracy, and real-time effectiveness, highlighting the need for an improved and more reliable solution.

IV. METHODOLOGY

i. Data Collection

The data collection phase involves gathering a dataset of phishing and legitimate URLs from reliable sources such as publicly available datasets and CSV files. The collected dataset contains labeled URLs, where each URL is classified as either phishing or legitimate. This data is essential for training and testing the machine learning model. The dataset includes various URL patterns that help the system learn the characteristics of malicious and safe websites. Proper data collection ensures that the model can accurately identify phishing websites in real-world scenarios.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503394|

ii. Feature Extraction

In this step, important features are extracted from the URLs to help the model understand their structure. These features include URL length, presence of HTTPS, number of special characters, domain information, and other suspicious patterns. The FeatureExtraction module processes each URL and converts it into a numerical feature vector. These features act as input for the machine learning model.

iii. Model Training

During model training, a machine learning algorithm such as Random Forest or Gradient Boosting is used to learn patterns from the dataset. The extracted features and their corresponding labels (phishing or legitimate) are used to train the model. The model learns to identify differences between safe and malicious URLs. After training, the model is saved using a pickle file for later use in prediction.

iv. URL Validation

Before making predictions, the input URL is validated to ensure it is correct and usable. This includes checking the URL format using validators, verifying domain existence using socket, and checking website availability using HTTP requests. This step helps prevent incorrect predictions for invalid or non-existent URLs, improving system reliability.

v. Prediction & Classification

In this step, the validated URL is processed through feature extraction and passed to the trained machine learning model. The model predicts whether the URL is safe or phishing and provides a probability score. This classification helps users understand the level of risk associated with the website.

vi. Result Display

The final result is displayed on a Flask-based web interface. The system shows the safety percentage along with a clear message indicating whether the website is safe or unsafe. It also displays appropriate error messages such as "Invalid URL" or "No Website Found" when necessary. This step ensures that users can easily understand the output.

V. SYSTEM ARCHITECTURE

System architecture defines the overall structure of the phishing website detection system and explains how different components interact to perform the required tasks. It provides a high-level view of how data flows through various modules including input processing, validation, machine learning prediction, and result display. A well-designed architecture improves system performance, accuracy, scalability, and reliability.

This architectural design supports modular development, making it easy to update or modify individual components such as feature extraction or validation without affecting the entire system. It also ensures efficient processing of user input and provides accurate real-time results.

The architecture of the proposed system begins with the User Interface Layer, where users enter a URL through a Flask-based web application. This input is then passed to the Input Processing Layer, where basic preprocessing is performed, such as adding “https://” if missing and cleaning the URL.

Next, the system moves to the Validation Layer, where multiple checks are applied to ensure the correctness of the URL. This includes URL format validation using validators, domain existence verification using socket programming, and website availability checking using HTTP requests. If the URL is invalid or the website does not exist, appropriate messages are generated without passing the input to the model.

If the URL passes validation, it is forwarded to the Feature Extraction Layer, where important URL-based features such as length, HTTPS presence, domain characteristics, and special symbols are extracted and converted into numerical form.

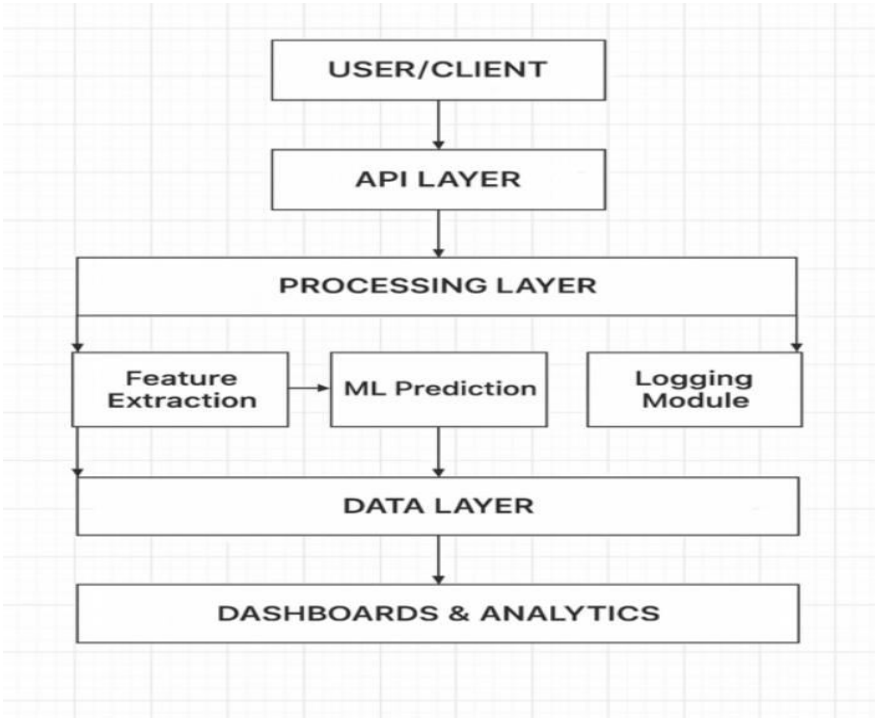
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

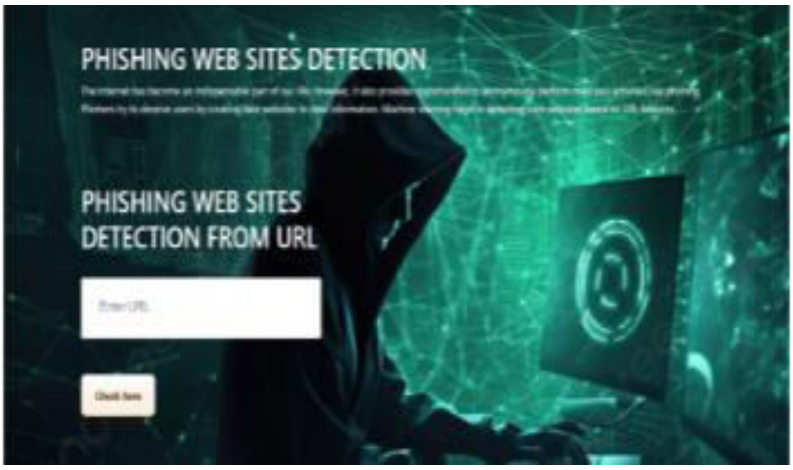
|DOI: 10.15680/IJIRSET.2026.1503394|



These features are then sent to the Machine Learning Layer, where a pre-trained model (such as Gradient Boosting or Random Forest) analyzes the input and predicts whether the website is phishing or legitimate. The model also provides a probability score indicating the confidence of the prediction.Finally, the output is sent to the Result Display Layer, where the Flask web interface presents the results to the user in a clear and understandable format. This includes safety percentage, classification (safe or unsafe), and error messages such as "Invalid URL" or "No Website Found."

The system architecture ensures smooth integration of validation techniques and machine learning, resulting in improved accuracy, reduced false predictions, and enhanced user trust. It provides a practical and efficient solution for detecting phishing websites in real-time environments.

VI. RESULTS AND OUTPUT



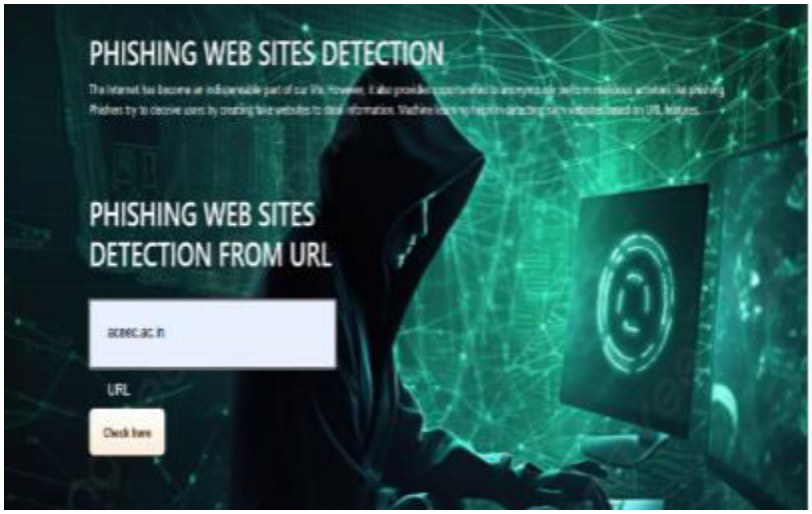
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

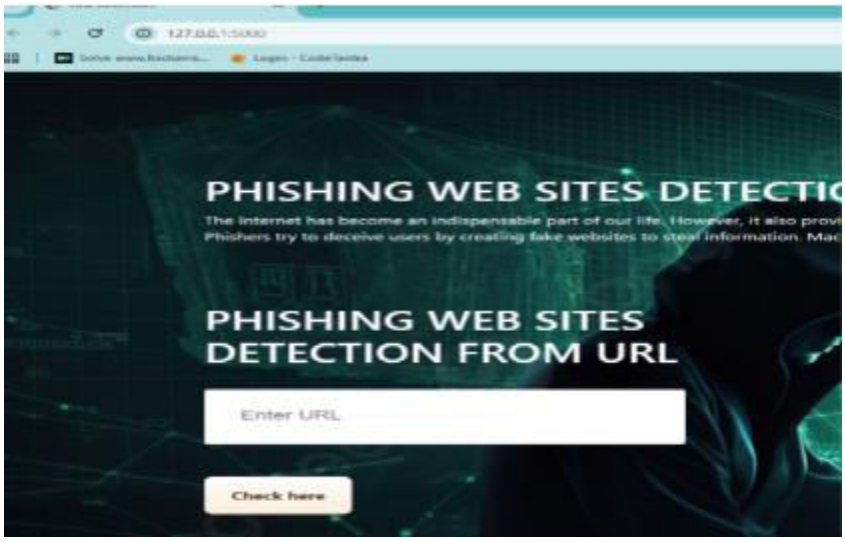
Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503394|





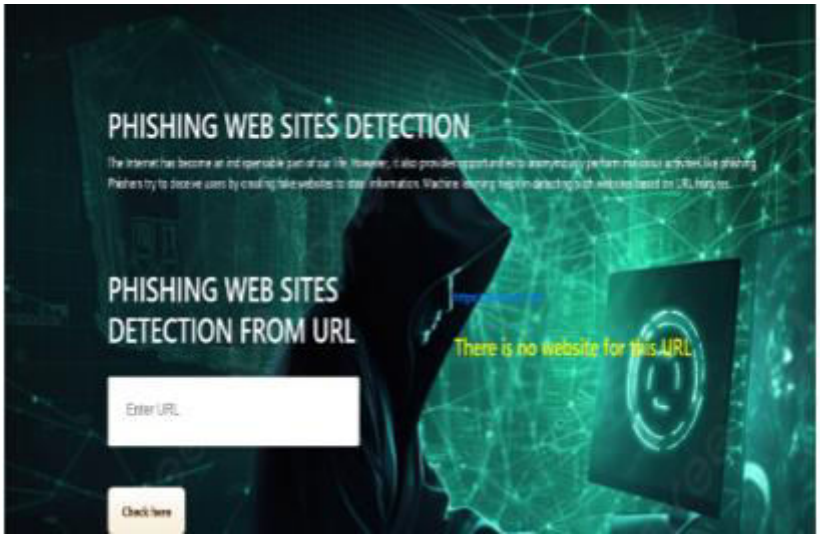
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503394|



VII. CONCLUSION

The rapid growth of internet usage and digital services has significantly increased the number of phishing attacks, making website security a major concern for users. Traditional rule-based detection systems are no longer sufficient to identify evolving phishing techniques, as they often fail to detect new or sophisticated malicious URLs. To address this challenge, this project developed a Machine Learning-based phishing website detection system that combines intelligent prediction with real-time URL validation to improve both accuracy and reliability.

The main objective of this project was to design a system capable of accurately classifying websites as legitimate or phishing while providing clear and understandable results to users. The system uses a machine learning model trained on URL-based features such as URL length, HTTPS presence, domain characteristics, and special symbols. To enhance system reliability, additional validation mechanisms were integrated, including URL format checking, domain existence verification, and website availability testing.

Handling invalid and non-existent URLs was an important aspect of this project, as traditional systems often produce incorrect predictions for such inputs. By incorporating validation techniques along with feature extraction and model prediction, the system ensures more accurate and meaningful results. Evaluation of the system shows that it can effectively distinguish between safe and phishing websites while reducing false predictions and improving user confidence.

Overall, the proposed system demonstrates that combining machine learning with validation techniques can significantly improve phishing detection performance. It provides a reliable and user-friendly solution for identifying malicious websites in real-time environments. This project also offers a strong foundation for future enhancements, such as integrating advanced models, real-time threat intelligence, and improved feature analysis for more robust cybersecurity applications.

VIII. FUTURE SCOPE

Future enhancements of the Phishing Website Detection System can improve its accuracy, scalability, and real-world applicability. One major improvement is the implementation of real-time phishing detection, where URLs are analyzed instantly as users browse the internet. This would help in identifying malicious websites immediately and preventing users from accessing harmful content, thereby enhancing cybersecurity and reducing risks.

The system can be further enhanced by integrating it with browser extensions or web security tools, allowing automatic detection of phishing websites without requiring manual input from users. Additionally, the project can be expanded by incorporating more advanced machine learning and deep learning models, which can improve detection accuracy and

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 3, March 2026

|DOI: 10.15680/IJIRSET.2026.1503394|

adapt to evolving phishing techniques. Continuous learning mechanisms can also be introduced so that the system updates itself based on new phishing patterns and real-time data.

Another possible enhancement is the integration of real-time threat intelligence and blacklist databases to strengthen detection capabilities. The system can also be deployed as a cloud-based application or mobile application, making it accessible to a wider range of users. Furthermore, advanced features such as user authentication, role-based access, and security monitoring dashboards can be added for administrative control and better system management.

Some possible future improvements include:
• Real-time URL monitoring and instant phishing alerts
• Integration with web browsers and security tools
• Use of machine learning and deep learning models• Continuous model training with updated phishing atasets
• Deployment as a cloud-based or mobile application
• Integration with cybersecurity threat intelligence systems

REFERENCES

1. Sahingoz, O. K., Buber, E., Demir, O., & Diri, B. (2019).
Machine Learning Based Phishing Detection from URLs.
Expert Systems with Applications.
https://doi.org/10.1016/j.eswa.2019.01.039
2. Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009).
Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs.
*Proceedings of the 15th ACM SIGKDD Conference.*https://doi.org/10.1145/1557019.1557153
3. Phishing Websites Dataset.
UCI Machine Learning Repository.
https://archive.ics.uci.edu/ml/datasets/phishing+websites
4. Pedregosa, F., et al. (2011).
Scikit-learn: Machine Learning in Python.
Journal of Machine Learning Research .
https://jmlr.org/papers/v12/pedregosa11a.html
5. Flask Documentation.
Python Web Framework.
https://flask.palletsprojects.com/

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details