International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 12, Issue 6, November-December 2025

Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 6, November-December 2025 ||

DOI:10.15680/IJARETY.2025.1206028

Intelligent Phishing URL Detection using Data Science

Shravani R[1], Pooja Taragar[2]

PG Student, Dept. of MCA, City Engineering College, Bengaluru, India[1]

Assistant Professor, Dept. of MCA, City Engineering College, Bengaluru, India[2]

ABSTRACT: In recent years, the Internet has become an essential part of our daily lives. 5.44 Social media is used by billions of people and the Internet worldwide, with more than 90% of them using social media. In the past ten years, several incidents have raised The necessity of digital education, commerce, and employment, notably COVID-19, which has accelerated the use of digital services. However, the security of publicly available data remains a serious issue. Network security is a concept, method, and approach that has been in use as long as networks. As long as information is shared, attacks, theft, and fraud attempts will persist. This article looks at many strategies to reduce the exploitation of personal information as well as defences against such risks. The newly created dataset was subjected to a Random Forest classifier following the merging of several datasets. There have been efforts to enhance the dataset and increase the model's accuracy, even if the precision of the linked models is good enough. The freshly proposed dataset was used to achieve accuracy. Artificial intelligence is crucial for strengthening cybersecurity defences, but it also facilitates hacks. Because of its dual nature, artificial intelligence can be employed for both offensive and defensive purposes in the cyberspace.

KEYWORDS: phishing, artificial intelligence, Cyber security, Phishing attack, Machine learning, Classification algorithms, Cyber-attack detection.

I. INTRODUCTION

One of the most prevalent and dangerous cyberattacks, when hackers exploit usernames, passwords, bank account details, and sensitive information. In phishing assaults, malicious people fabricate phony URLs and websites that closely mimic authentic ones in an attempt to trick users into believing them. The rapid growth of online services, e-commerce, and digital banking has led to an increase in the frequency and sophistication of phishing attempts. Blacklists, rule-based systems, and human analysis are the mainstays of traditional phishing detection methods. Despite being quick and easy, these techniques have significant drawbacks. Zero-day phishing URLs are newly constructed malicious links that haven't been published, and blacklist-based methods are unable to identify them. Additionally, rule-based systems frequently produce high false-positive rates and find it difficult to adjust to changing attack patterns.

Using statistical analysis and machine learning, Data-Based Intelligent Phishing URL Identification Science circumvents these limitations by automatically identifying phishing URLs based on characteristics and patterns extracted from the URLs themselves. Instead of depending solely on known malicious links, data-driven algorithms can generalise to new URLs by discovering hidden relationships from prior data.

Because they may function in real-time settings, intelligent phishing detection systems are appropriate for network security applications, email gateways, and browsers. These systems can adjust to new phishing tactics and offer better defense against zero-day assaults by continuously learning from fresh data and feedback.

In conclusion, data science-based intelligent phishing URL detection offers a scalable, flexible, and effective defense against contemporary phishing attacks.

II. LITERATURE SURVEY

1. **Title**: Machine Learning Methods for Phishing URL Identification
Authors: A. Kumar, R. Sharma
Abstract: Malicious URLs are used in phishing attempts to trick users and obtain private data. This research describes a machine learning-based method that uses lexical information taken from URLs to identify phishing URLs. To

differentiate phishing URLs from authentic ones, classifiers like Support Vector Machines and Random Forests are learned using datasets with labels. Results from experiments indicate that machine learning methods outperform traditional blacklist-based solutions in locating fresh phishing URLs.

2. **Title**: A Data Science How to Spot Phishing Websites
Authors: S. Patel, K. Mehta
Abstract: Phishing websites mimic reputable websites in order to deceive visitors into divulging personal information. A data science-based phishing detection model utilising host-based and URL-based information is presented in this work. Accuracy and recall are utilized to gauge The effectiveness of several classifications systems. The results indicate that in contrast to rule-based systems, data-driven methods offer better detection accuracy and flexibility.

3. **Title**: Zero-Day Phishing URL Detection Using Anomaly Detection
Authors: L. Wang, H. Li
Abstract: Zero-day phishing URLs cannot be detected by traditional signature-based techniques. This paper presents an anomaly detection framework that mimics normal URL behaviour and identifies deviations as potential phishing attempts. To find dubious URLs, one-class SVM and autoencoder models are used.

4. **Title**: Character-Level Deep Learning-Based Phishing Detection URLs
Authors: J. Smith, M. Johnson
Abstract: Complex URL obfuscation techniques are used in modern phishing assaults, making them challenging to identify with handmade features. A character-level deep learning model that gains knowledge directly from raw URLs is presented in this research. The program achieves excellent detection accuracy by automatically identifying structural patterns in URLs. The efficacy of deep learning methods in lowering false positives is confirmed by experimental examination.

5. **Title**: An Intelligent Hybrid System for Real-Time Phishing URL Detection
Authors: P. Rao, N. Gupta
Abstract: An intelligent hybrid detection technique for phishing URLs that combines anomaly detection and supervised machine learning techniques is presented in this paper. Ensemble classifiers are used to analyse lexical and statistical data that are gathered from URLs. The hybrid method makes it possible to identify zero-day phishing URLs and increases detection accuracy. The technique can be implemented in real-time for email filtering and browser security applications.

III. METHODOLOGY

Existing Problem:
Phishing attempts, which attempt to obtain private information such as bank account details, login passwords, and personal information, continue to be a major concern due to malicious URLs. The majority of systems that detect fraudulent emails utilize signatures, rules, or blacklists. These techniques are easy to apply, However, they don't work against zero-day phishing URLs—new dangerous links that haven't been submitted to databases yet.

Because attackers often change their URL structures to avoid being caught, rule-based systems are hard to keep and update because the rules and thresholds have to be set by hand. Updates to blacklist-based methods are also slow, and they can't offer proactive security. Also, many of the systems that are already in place have high false-positive rates, which means they mistakenly label legal URLs as phishing, which makes users less likely to trust them.

Proposed Solution:
To overcome the shortcomings of traditional phishing detection methods, a data science-based intelligent phishing URL detection system is proposed. The proposed approach use machine learning techniques to automatically assess URLs and classify them as authentic or phishing based on learnt patterns, as opposed to static rules or blacklists. Initially, the system collects URLs from various sources, such as emails, user input, and browser queries. These URLs are then pre-processed and standardised to ensure uniformity.

Proposed Methodology:

In order to reliably identify URLs such as phishing or real, the suggested methodology for intelligent identification of phishing URLs utilizing data science employs a methodical workflow. The approach is made to be scalable, effective, and able to identify zero-day phishing URLs.

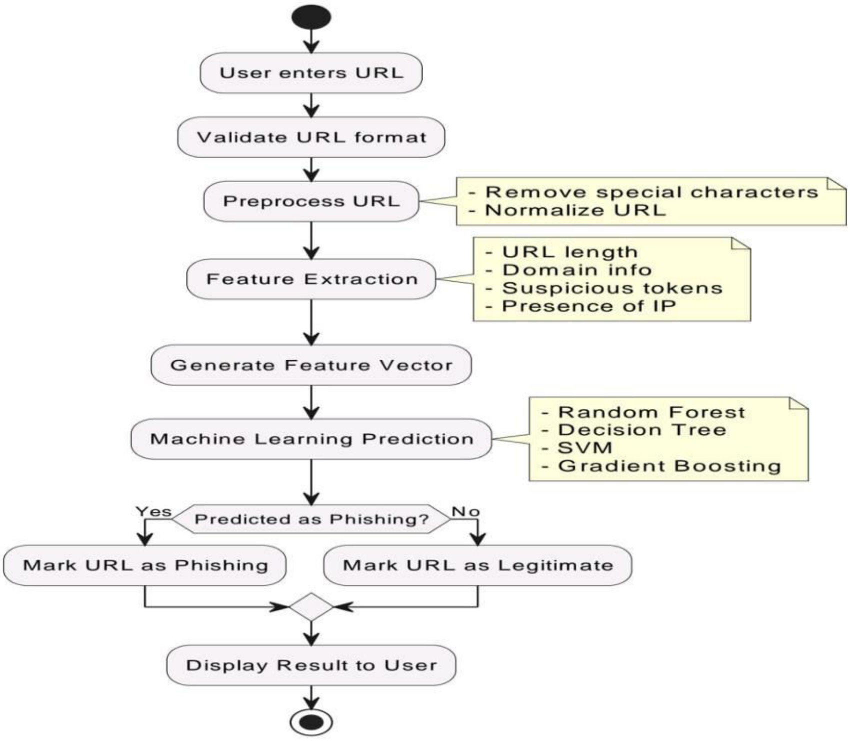

Fig 1: Proposed Methodology

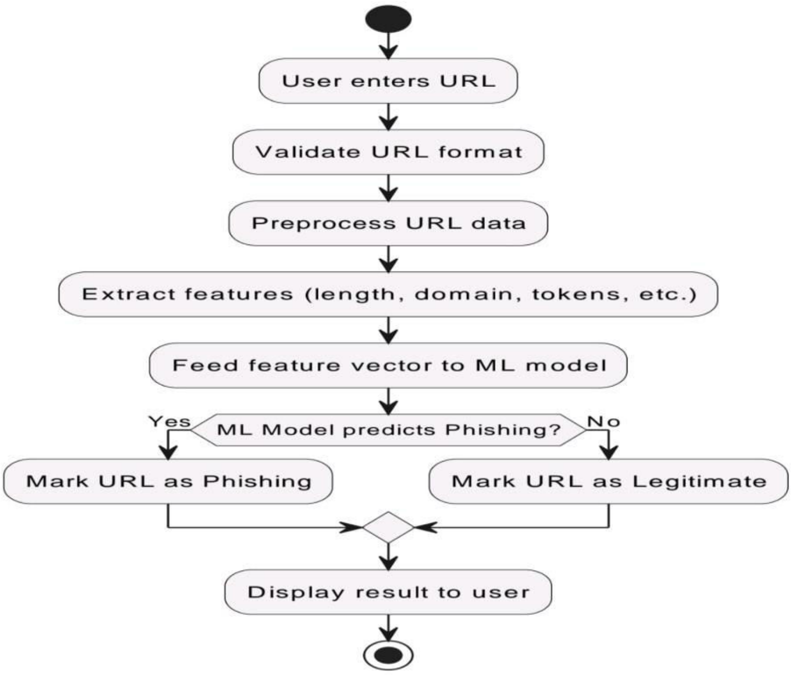

Fig 2: Activity Diagram

IV. SYSTEM DESIGN

The system design for data science-based intelligent phishing URL detection consists of several of linked modules that cooperate to give precise and instantaneous detection. Users or programs can submit URLs through the User Interface module, and the detection results can be shown as safe notifications or phishing alerts. The URL Collection module collects URLs for analysis from several sources, including emails, user input, and browser activity. To ensure consistency, the Preprocessing module decodes characters and eliminates superfluous symbols from the gathered URLs. While the Feature Selection module selects the most pertinent aspects to increase efficiency and accuracy, the Feature Extraction module examines the URLs and extracts significant lexical and statistical characteristics that point to phishing behavior.

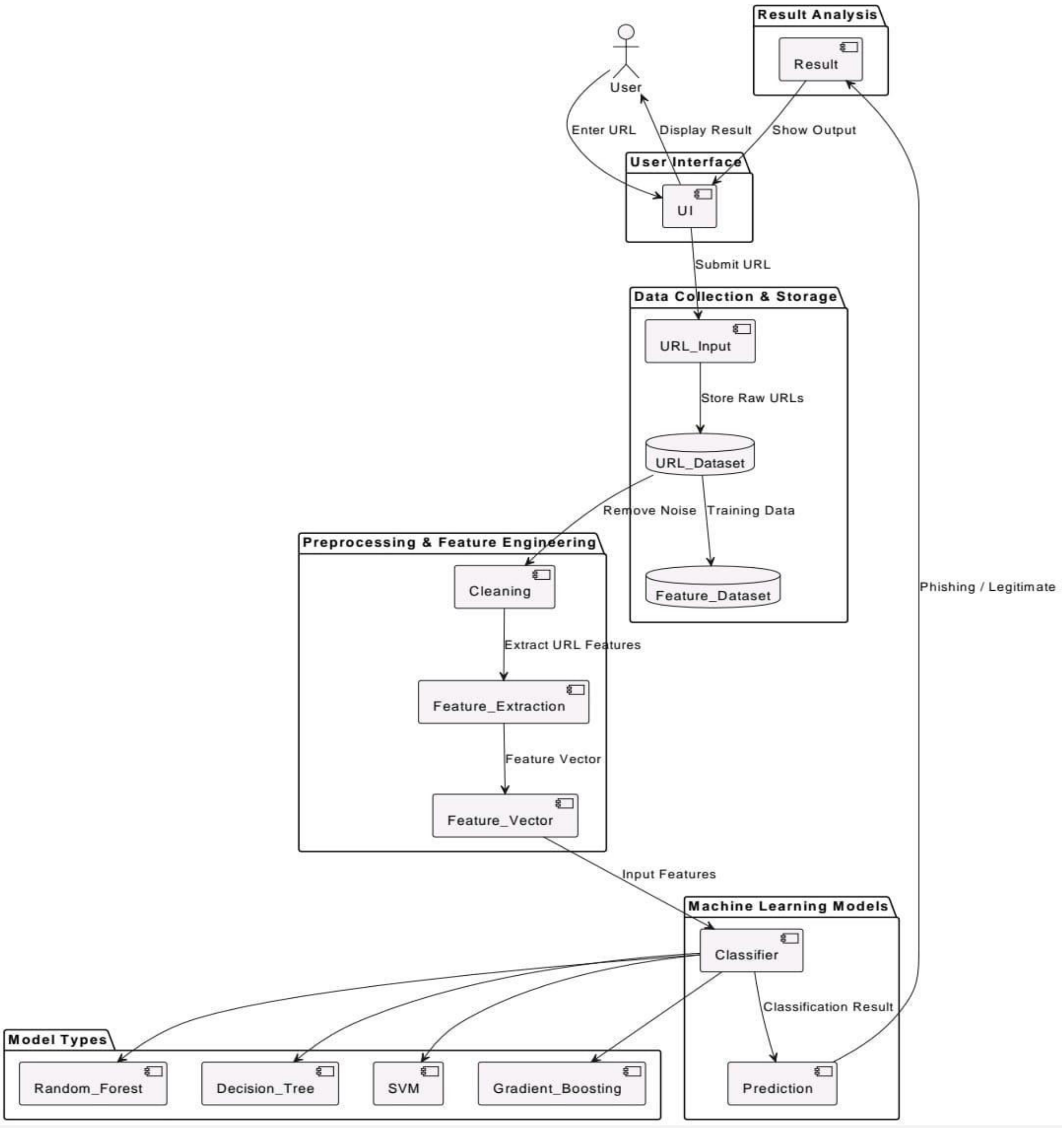

Fig 3: System design

V. SYSTEM ARCHITECTURE & DESIGN

To guarantee effectiveness, scalability, and real-time operation, the system architecture and design for intelligent phishing URL detection utilizing data science are arranged as a modular, layered framework. The User Interface module, which enables users or programs to provide URLs and get safe or phishing alerts, is where the architecture starts. After gathering URLs from numerous sources from numerous sources, such as, including emails, web traffic, and browsers, the URL Collection module sends them to the Preprocessing module for cleaning, decoding, and normalization. The Feature Selection module filters the most pertinent aspects to improve performance after the Feature Extraction module extracts important lexical in addition to statistics data from the URLs. The Engine for The field of machine learning is made up of trained classifiers that can recognize phishing trends, processes these features. The design includes a Zero-Day Detection module that uses anomaly detection or hybrid learning approaches to increase resistance against new threats.

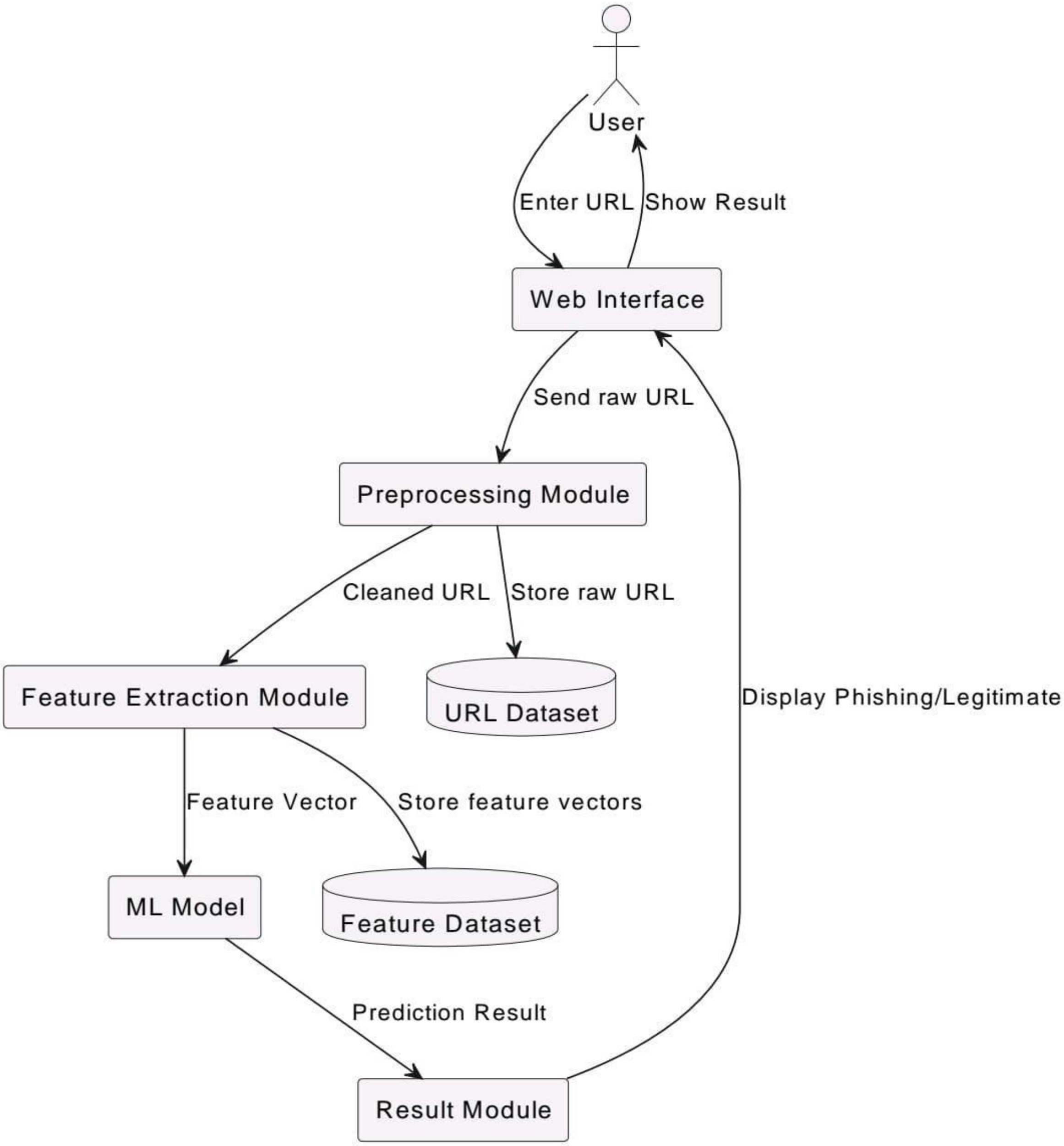

Fig 4: Data flow diagram

VI. IMPLIMENTATION

The first step in implementing the data science-based clever method for identifying fraudulent URLs is assembling a dataset from reputable sources that includes both real and phishing URLs. In order to preserve a uniform format, the URLs first go through a preprocessing step where they are cleaned, decoded, and normalized. Following preprocessing, a feature extraction procedure is used to calculate significant lexical and statistical characteristics like the length of the URL, the amount of numbers the quantity of special characters, and whether IP addresses are present. These characteristics are arranged into a structured dataset that may be utilised for machine learning investigation following their extracted. A classification model is trained using the labelled dataset, like Random Forest or Support Vector Machine, to identify patterns that differentiate phishing URLs from authentic ones. The model is incorporated into the system to carry out real-time URL classification after it has been trained. Anomaly detection techniques or ensemble approaches that find odd URL patterns might be utilised in the implementation help improve the identification of zero-day phishing URLs. The user interface receives the categorization result from the system, which indicates if the URL is safe or phishing. Lastly, the model is frequently retrained using detected URLs and user feedback that are kept in a database, guaranteeing ongoing development and flexibility in response to changing phishing threats.

VII. RESULTS & DISCUSSION

The intelligent phishing URL The detection system was evaluated using a labelled dataset that had both authentic and phishing URLs. The experimental data shown the machine learning model's capacity to classify URLs with great precision and improved precision and recall values. Lexical and statistical features, like the length of the URL, The quantity of special characters, and the quantity of numbers, significantly contributed to accurate recognition. Most of phishing URLs, even A few that had never been observed before, were determined by the system, demonstrating its capacity to detect zero-day phishing attacks. Seldom were legitimate URLs wrongly classified as phishing due to the comparatively low false-positive rate. The results indicate that data science–based phishing URL detection is more effective than traditional blacklist and rule-based approaches. The model's capacity to learn patterns from data enables it to generalize well to new and evolving phishing techniques. The integration of anomaly detection further strengthens zero-day attack detection by identifying unusual URL structures. Additionally, the modular system design supports real-time implementation with minimal computational overhead. However, The quality and diversity of the training dataset determine the system's performance, and idea drift requires regular retraining. All things considered, the suggested method offers a scalable, flexible, and clever way to detect phishing URLs.

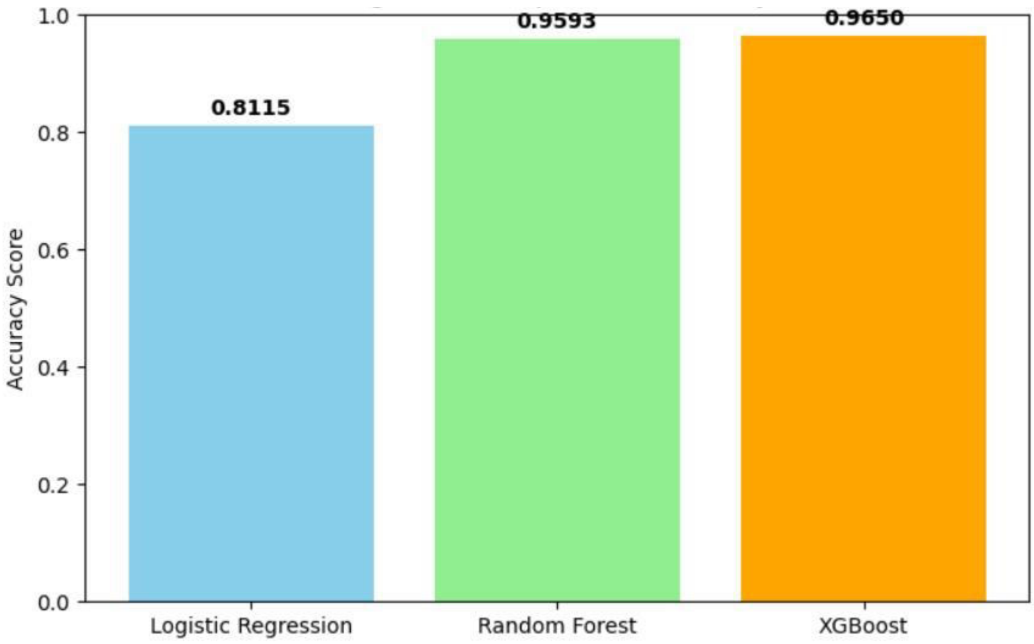

Fig: 5 algorithm comparison-Accuracy

VIII. CONCLUSION

In conclusion, the data science-based intelligent phishing URL detection system offers a proactive and successful defense against phishing attempts. The solution overcomes the drawbacks of conventional blacklist and rule-based approaches by utilizing machine learning techniques and URL-based feature analysis, especially in identifying zero-day phishing URLs. The suggested method shows excellent accuracy, fewer false positives, and flexibility in response to changing phishing tactics. Real-time deployment and ongoing development through feedback and model retraining are made possible by its modular design. By providing a scalable, intelligent, and dependable method for spotting bad URLs and shielding users from phishing attacks, this system improves cybersecurity overall.

IX. FUTURE ENHANCEMENTS

Future enhancements to the intelligent phishing URL detection system using data science can enhance its precision, resilience, and real-world applicability. sophisticated deep learning Models , like transformer-based architectures and character-level neural networks, can be integrated to automatically learn complex URL patterns without manual feature engineering. The system can also be enhanced by incorporating real-time threat intelligence feeds and domain reputation services to strengthen detection of emerging phishing campaigns. Implementing online and incremental learning techniques would allow the model to adapt continuously to new phishing behaviors without complete retraining. Additionally, integrating browser extensions or email gateway plugins can enable proactive phishing prevention at the user level. Explainable AI techniques can be added to provide transparent reasoning behind model decisions, improving user trust and analyst understanding. Finally, extending the system to analyze email content, webpage content, and network traffic alongside URLs would create a comprehensive multi-layer phishing detection framework.

REFERENCES

1. Korkmaz M, Sahingoz OK, Diri B (2020) “Detection of phishing web sites by using machine learning-based URL analysis.”
2. Jalil S, Usman M (2020) “A review of identifying fraudulent URLs using machine learning classifiers” Springer Adv Intell Syst Comput 1251: 646–665
3. El Aassal A, Baki S, Das A, Verma RM (2020)” An indepth bench Assessing and assessing research on phishing detection for security requirements. IEEE Access 8:22170–22192
4. Shahrivari V, Darabi MM, Izadi M (2020) “Machine learning algorithms for phishing detection”
arXiv 2009.11116.
5. Aburub F, Hadi W (2021) “A novel approach to phishing detection based on association categorisation “websites” J Theoret Appl Inf Technol 99(1):147–158
6. F. Feng, Q. Zhou, Z. Shen, X. Yang, L. Han, and J. Wang, "The application of a novel neural network in the detection of phishing websites" J. Ambient Intelligence Humanized Comput., vol. 15, no. 3, pp. 1865–1879, March 2024, doi: 10.1007/s12652-018-0786-3
7. "Improving phishing detection: A novel hybrid deep learning framework for cyber crime forensics," by F. S. Alsubaei, A. A. Almazroi, and N. Ayub 2024; doi: 10.1109/ACCESS.2024.3351946; IEEE Access, vol. 12, pp. 8373–8389.

IJARETY

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975

Impact Factor: 8.152

www.ijarety.in

editor.ijarety@gmail.com