International Journal of Engineering & Extended Technologies Research (IJEETR)

|ISSN: 2322-0163| www.ijeetr.com | A Bimonthly, Peer Reviewed, Scholarly Indexed Journal |

| Volume 8, Issue 2, March - April 2026 |

DOI:10.15662/IJEETR.2026.0802010

ML-Based Phishing Website Detection & Prevention System

Dr.S.Kavitha M.Sc., M.Phil, Ph.D. [1] , S. AKALYA [2]

Head of the Department, Department of Computer Science, Sakthi College of Arts and Science for Women, Oddanchatram, Tamilnadu, India[1]

M. Sc (Computer Science), Department of Computer Science, Sakthi College of Arts and Science for Women, Oddanchatram, Tamilnadu, India[2]

Publication History: Received: 23.02.2026; Revised: 10.03.2026; Accepted: 13.03. 2026; Published: 18.03.2026

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 Detection, Cybersecurity, Machine Learning, URL Analysis, Feature Extraction, Random Forest, Support Vector Machine (SVM), Decision Tree, Website Security, Malicious URL Detection, Fraud Prevention, Network Security

I. INTRODUCTION

Nowadays Phishing becomes a main area of concern for security researchers because it is not difficult to create the fake website which looks so close to legitimate website. Experts can identify fake websites but not all the users can identify the fake website and such users become the victim of phishing attack. Main aim of the attacker is to steal banks account credentials. In United States businesses, there is a loss of US$2billion per year because their clients become victim to phishing. In 3rd Microsoft Computing Safer Index Report released in February 2014, it was estimated that the annual worldwide impact of phishing could be as high as $5 billion. Phishing attacks are becoming successful because lack of user awareness. Since phishing attack exploits the weaknesses found in users, it is very difficult to mitigate them but it is very important to enhance phishing detection techniques.

II. LITERATURE SURVAY

1. Mahajan (2018) conducted one of the early studies applying machine learning for phishing detection using algorithms such as Support Vector Machine (SVM), Random Forest, and Decision Trees. The research demonstrated that machine learning models could effectively identify phishing websites by learning from URL features, achieving high accuracy with minimal false positives. Similarly, Kiruthiga and Akila (2019) examined multiple algorithms and found Random Forest and Decision Tree to be efficient classifiers for URL-based detection.

2. Kulkarni and Brown (2019) emphasized the significance of using feature engineering and data preprocessing to improve model performance. Their study, published in IJACSA, highlighted that combining URL, HTML, and network-based features enhanced prediction reliability.

3. Safi and Singh (2023) provided a systematic literature review summarizing phishing detection approaches and challenges. They discussed how feature extraction, dataset selection, and evaluation metrics significantly influence accuracy and model robustness. Mahmoud Khonji, Iraqi, and Jones conducted a comprehensive survey analyzing literature trends, categorizing techniques into blacklist-based, heuristic-based, and ML-based methods, and stressing the importance of standard benchmark datasets.

4. Qasim (n.d.) presented a detailed review of ML techniques, outlining how supervised learning dominates phishing detection but unsupervised and hybrid approaches are gaining importance. Likewise, Arathi Krishna et al. (IJERT) surveyed URL-based ML systems and concluded that ensemble classifiers outperform single models in large datasets.

5. Padmini and Usha Sree (2024) explored feature selection techniques for URL-based detection and demonstrated that selecting significant lexical and domain features improves prediction speed and accuracy. Yang, Zhao, and Zeng (2019) supported this by integrating feature selection with ML algorithms, achieving a balanced trade-off between accuracy and model complexity.

6. Phanindra Kumar et al. created a documentation-based model focusing on URL and content features extracted via Python, achieving efficient detection through hybridized algorithms. Similarly, Garje et al. (IJCRTI020051.pdf) emphasized lightweight URL analysis for faster phishing identification.
7. Dutta (2021) applied supervised learning models on PLoS ONE datasets, showcasing that SVM and Logistic Regression performed exceptionally well for phishing detection. Alazaidah et al. (n.d.) at the American Academic Research University combined heuristic and ML techniques to develop a hybrid system integrating URL, HTML, and host-based attributes.

8. Zieni, Massari, and Calzarossa (2023) introduced advanced ML and ensemble models such as Gradient Boosting and XGBoost in phishing detection. They concluded that ensemble learning significantly reduces false detection rates and enhances generalization on unseen data. Tang and Mahmoud (n.d.) also utilized multi-layer ML techniques, improving accuracy using combined URL and page content features.

9. Alzboon (n.d.) presented a study titled Guardians of the Web, demonstrating the integration of deep learning with ML classifiers to counter new phishing tactics dynamically. Das Guptta et al. (2024) similarly analyzed diverse machine learning models in phishing detection using an Emerald publication, concluding that hybrid deep learning provides resilience against evolving phishing strategies.

10. Kulkarni (A.D., n.d.) replicated ML-based detection frameworks using different datasets, confirming that algorithms like Random Forest and Naïve Bayes produce consistent accuracy across domains. Bhosale et al. (n.d.) proposed a phishing detection model using supervised learning and URL-based datasets, focusing on practical implementation within browser security layers

III. THEORETICAL BACKGROUND

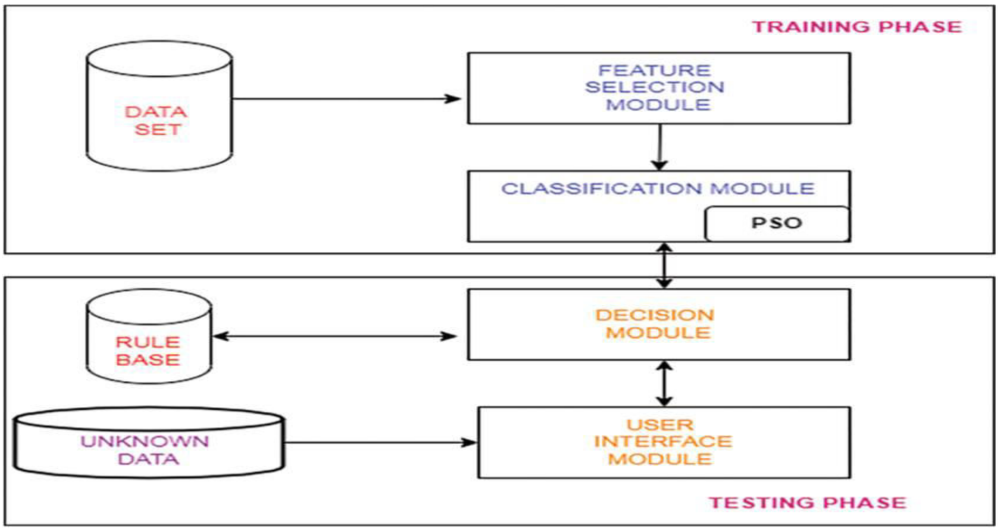
3.1 PROBLEM IDENTIFICATION

- Anti-phishing strategies involve educating netizens and technical defense. In this paper, we mainly review the technical defense methodologies proposed in recent years. Identifying the phishing website is an efficient method in the whole process of deceiving user informationAlong with the development of machine learning techniques, various machine learning based methodologies have emerged for recognizing phishing websites to increase the performance of predictions.The primary purpose of this paper is to survey effective methods to prevent phishing attacks in a real-time environment

3.2 PROBLEM SOLVING

- The most frequent type of phishing assault, in which a cybercriminal impersonates a well-known institution, domain, or organization to acquire sensitive personal information from the victim, such as login credentials, passwords, bank account information, credit card information, and so on.Emails containing malicious URLs in this sort of phishing email contain a lot of personalization information about the potential victim.To spear phish a "whale," here a top-level executive such as CEO, this sort of phishing targets corporate leaders such as CEOs and top-level management employeesTo infect the target, the fraudster or cyber-criminal employs a URL link.

3.3 SYSTEM ARCHITECTURE



IV. SYSTEM IMPLEMETATION

4.1. MODULE:

- User Interface Module
- URL Feature Extraction Module
- Machine Learning Training Module
- Prediction & Classification Module
- Prevention / Blocking Module
- Admin Dashboard Module

4.2 MODULE DESCRIPTION:

1. User Interface Module

- The user enters a website link (URL).
- Simple input form that sends the URL to the backend.
- Displays the final classification result.

2. URL Feature Extraction Module

Extracts important features such as:

- URL length
- Presence of "@" symbol
- Number of subdomains
- HTTPS / SSL certificate validity
- Domain age
- Redirections
- "https" spoofing
- Suspicious keywords

3. Machine Learning Training Module

- Preprocessing & cleaning dataset
- Splitting dataset into training/testing
- Models used: Random Forest, Decision Tree, SVM, KNN
- Training the model on thousands of legitimate and phishing URLs
- Saving the trained model (.pkl file)

4. Prediction & Classification Module

- The extracted features of the user-provided URL are passed to the trained ML model.
- Output will be:

1. Legitimate
2. Suspicious
3. Phishing

5. Prevention / Blocking Module

When a phishing URL is detected:

- The system alerts the user
- The page is automatically blocked
- A warning message is shown (similar to Chrome "Deceptive Site Ahead")

6. Admin Dashboard Module

- Displays statistics
- Accuracy reports
- Dataset management
- ML model re-training option

V CONCLUSION

The increasing popularity of phishing websites stands as a significant and evolving threat within the digital domain. These platforms are particularly designed to mislead users, give in sensitive information, and propose substantial risks to cybersecurity infrastructure. Considering the scope of these dangers, it is essential to take the detection of phishing websites very seriously. This study has investigated phishing detection in great detail, evaluating the value and efficacy of a wide range of DL and ML models. By comparing the performance of several models, such as SVM, DT, RF, KNN, GRU, LSTM, RNN, and ensemble learning models like XGBoost, AdaBoost, and RF, the study established a distinction between authentic and phishing domains. Among these many models, the ensemble learning strategy that is, RF in particular has proven quite effective, with 99% accuracy.

REFERENCES

1. Howe, A. von Mayrhauser, and Mraz, R. T. Test case generation as an AI planning problem. Automated Software Engineering, 4:77-106, 1997.
2. Koehler, J., Nebel, B., Hoffman, J., and Dimopoulos, Y. Extending planning graphs to an ADL subset. Lecture Notes in Computer Science, 1348:273, 1997.
3. Treutner, M. F., and Ostermann, H. Evolution of Standard Web Shop Software Systems: A Review and Analysis of Literature and Market Surveys.
4. CS-Cart.com (Simbirsk Technologies Ltd), © 2004-2013.http://www.cs-cart.com/
5.Ofbiz, the Apache Open for Business Project. Retrieved on 2013."http://ofbiz.apache.org/index.html"
6.Comparison of shopping cart software. Retrieved on June 28, 2013.
http://en.wikipedia.org/wiki/Comparison_of_shopping_cart_software
7.Demonstrating how the web server Operates using PHP5/24/2018
8.All about frontend controls in php http://www.msdn.microsoft.com/
9.Wikipedia for various diagrams & testing methods http://www.wikipedia.org/
10.Cool text for Images and Buttons http://cooltext.com/
11.K-State Research Exchange for samples in report writing http://krex.k-state.edu/dspace/handle/2097/959
12. Smart Draw for drawing all the Diagrams used in this report. http://www.smartdraw.com/
13. Sample Ecommerce Application http://www.NewEgg.com
14. Ajax Toolkit controls http://asp.net/ajax