International Journal of Scientific Research & Engineering Trends
Volume 10, Issue 6, Nov-Dec-2024, ISSN (Online): 2395-566X

Smart Shields against Cyber Threats: Machine Learning-Driven Phishing URL Detection

Syeda Aynul Karim, Md. Juniadul Islam, Ishtiaq Hoque Farabi
Department of Computer Science,
American International University-Bangladesh (AIUB), Dhaka, Bangladesh

Abstract- Phishing attacks remain a prevalent cybersecurity threat, exploiting vulnerabilities in digital platforms to compromise sensitive user data. This paper introduces a novel machine learning-based framework for phishing URL detection, combining advanced feature engineering techniques and classification algorithms. By integrating lexical attributes, WHOIS data, and ranking metrics like PageRank and Alexa Rank, our approach enhances detection accuracy and minimizes false positives. Experimental results demonstrate superior performance across classifiers, achieving an accuracy of 99.8% using Support Vector Machines. The framework's modular design ensures adaptability to evolving phishing tactics and scalability for enterprise deployment. This research lays the foundation for future advancements in AI-driven cybersecurity solutions.

Index Terms- Phishing Detection, Machine Learning, Cybersecurity, URL Analysis, Feature Engineering,

I. INTRODUCTION

Phishing attacks are among the most pervasive and damaging forms of cybercrime, targeting individuals and organisations by deceiving them into revealing sensitive information. These attacks often exploit trust in legitimate platforms by mimicking their URLs and interfaces to harvest data such as usernames, passwords, financial credentials, and personal identification. As the digital landscape grows more complex, the sophistication of phishing schemes has also increased, making them harder to detect and prevent. Phishing attacks not only result in significant financial losses and data breaches but also erode trust in digital platforms, particularly in critical domains such as online banking, e-commerce, and government services [1].The rapid evolution of phishing techniques has outpaced traditional detection systems, exposing critical vulnerabilities in existing security measures. Legacy approaches such as blacklist-based systems, heuristic rule engines, and manual inspections often fail to address new and advanced phishing tactics. These methods frequently suffer from high false positive rates and limited scalability, which diminishes their effectiveness in real-world applications.In response to these challenges, machine learning (ML) has emerged as a promising tool for phishing detection, leveraging patterns and features extracted from large datasets to improve detection accuracy [2]. However, many current ML-based solutions are hindered by narrow feature sets, an overreliance on specific datasets, and limited adaptability to novel phishing strategies.This research aims to bridge these gaps by introducing a robust, machine learning-based detection framework that leverages multidimensional feature analysis to enhance the accuracy and scalability of phishing detection systems [3]. Unlike traditional approaches, this framework integrates diverse types of data attributes, including lexical characteristics, WHOIS data, and reputation-based metrics such as PageRank and Alexa Rank [1]. These attributes collectively capture a comprehensive view of phishing behaviors, enabling the detection of even the most sophisticated phishing attempts.

Moreover, this study rigorously evaluates the effectiveness of various state-of-the-art classifiers, including Support Vector Machines (SVM), Random Forests, and Naive Bayes, to determine the most suitable algorithm for detecting phishing URLs [1]. Each classifier is assessed for its ability to process multidimensional features efficiently and provide accurate classifications. The resulting detection framework is designed to be both scalable and modular, ensuring its applicability across diverse datasets and evolving phishing scenarios. In addition to improving detection accuracy and minimizing false positives, the framework presented in this research is designed to address the real-world needs of cybersecurity[4]. Its modularity ensures easy integration into existing security infrastructures, while its adaptability to emerging threats makes it a future-proof solution. By focusing on multidimensional feature engineering and robust classification techniques, this work lays a strong foundation for the development of AI-driven phishing detection systems that can keep pace with the ever-changing tactics of cybercriminals. In the following sections, we provide a detailed account of the methodology, experimental evaluations, and the results achieved through this novel framework, highlighting its potential to redefine the landscape of phishing detection and cybersecurity.

II. RELATED WORK

The rise of internet usage has been accompanied by a surge in cybercrimes, particularly phishing attacks. Phishing, a prevalent form of cybercrime, manipulates users into revealing sensitive information through fraudulent websites. As the sophistication of these attacks has grown, researchers have explored numerous techniques for detecting phishing URLs using machine learning and associated methodologies [5]. This review critically examines key contributions in the field, focusing on the evolution of feature-based detection methods, the integration of machine learning models, and emerging challenges.

One of the earliest studies in phishing URL detection was conducted by Carolin and Rajsingh, who proposed a model utilizing association rule mining. This approach focused on extracting significant patterns from URL datasets, offering insights into how specific features distinguish between legitimate and malicious URLs [6]. Their work highlighted the importance of data mining techniques in understanding the structural and semantic characteristics of URLs, setting the stage for subsequent advancements in the domain [7].

Garera et al. advanced the field by introducing a framework that incorporated page ranking, domain presence in trusted tables, and lexical analysis to detect phishing URLs [3]. Their model emphasized the importance of combining multiple features to improve classification accuracy [8]. By integrating domain-based features with linguistic patterns, this study demonstrated the potential of feature fusion in mitigating the limitations of single-feature-based approaches.

In another significant contribution, Mohammed et al. leveraged Microsoft Reputation Services to develop a machine learning model for phishing detection [4]. Their research demonstrated the efficacy of combining proprietary reputation metrics with traditional URL features. This hybrid approach not only enhanced the detection of known phishing websites but also improved the system's adaptability to new threats.

Blum et al. presented a novel perspective by proposing a URL-based phishing detection model that operated independently of web page content [5]. Their work underscored the utility of lexical features, such as domain name structure and the occurrence of special characters, in identifying phishing attempts. This model's reliance on easily extractable attributes made it particularly appealing for real-time applications.

The CANTINA framework introduced a pioneering method for phishing detection based on HTML features. By analyzing the structural elements of web pages, CANTINA offered a robust solution for detecting phishing attempts that relied on visual and textual mimicry [6]. Subsequent enhancements to this framework incorporated additional features, further refining its performance.

Parekh et al. explored the use of document object model (DOM) features for phishing detection [7]. By leveraging the structural representation of HTML documents, their model captured subtle variations in webpage layouts that are indicative of phishing. This approach expanded the scope of feature-based detection to include visual and spatial attributes, bridging the gap between content-based and structural analysis.

Pradeepthi and Kannan extended the domain by proposing a visual detection method that analyzed text styles and images on webpages [8]. Their work demonstrated the importance of incorporating visual cues into phishing detection frameworks, particularly for attacks that relied heavily on user deception through graphical mimicry.

Fu's research on PhoneyC, a virtual honeypot for studying malicious URLs, provided critical insights into the behavior of phishing websites [9]. By simulating user interactions, PhoneyC identified dynamic features of phishing URLs, such as redirection patterns and script behaviors. This study highlighted the potential of behavioral analysis in complementing traditional static features.

Sahoo et al. introduced an innovative approach to phishing detection by converting webpages into images and analyzing their visual signatures. By leveraging techniques such as color histograms and spatial relationships, their model identified subtle differences between legitimate and phishing webpages. This visual analysis method demonstrated the effectiveness of integrating image processing techniques into phishing detection [10].

The growing complexity of phishing attacks has also spurred the development of machine learning models that incorporate advanced algorithms.

Decision trees and random forests have been widely adopted for their ability to handle high-dimensional datasets and provide interpretable results [11]. Light Gradient Boosting Machine (LightGBM) has emerged as a particularly promising technique, offering superior performance in terms of accuracy and scalability.

Logistic regression, a traditional statistical model, has been employed as a baseline in many studies. Its simplicity and interpretability make it a reliable choice for initial evaluations. However, its performance often lags behind ensemble methods like random forests and LightGBM in complex datasets [12].

Support Vector Machines (SVM) have been praised for their effectiveness in handling high-dimensional data. Their ability to model non-linear relationships using kernel functions makes them suitable for capturing intricate patterns in phishing URLs. However, their computational complexity can be a limitation in large-scale applications [13].

Naive Bayes, known for its simplicity and computational efficiency, has also been explored in phishing detection [14]. While it performs well on small datasets with clear feature separations, its assumption of feature independence can limit its applicability in more complex scenarios.

Despite these advancements, several challenges remain. The rapid evolution of phishing techniques necessitates continuous updates to detection frameworks [15]. The reliance on static features makes some models vulnerable to evasion tactics, such as the use of obfuscated URLs or trusted domains in phishing attacks. Additionally, the scarcity of labeled datasets poses a significant barrier to training robust machine learning models.

Our study builds upon the foundational work of previous researchers, integrating and advancing their methodologies to create a robust and adaptive phishing URL detection framework [16, 17].

The proposed model leverages a comprehensive set of features, including lexical, domain-based, and reputation-based attributes, combined with cutting-edge machine learning algorithms such as Light Gradient Boosting Machine (LightGBM), Random Forest, and Support Vector Machine (SVM).

By incorporating these advancements, our model not only addresses the limitations identified in prior studies but also sets a new benchmark for accuracy, scalability, and real-time applicability.

III. METHODOLOGY

This section outlines the step-by-step methodology used to develop and evaluate the machine learning-based phishing URL detection framework. The methodology is designed to ensure robustness, scalability, and high performance. A systematic flowchart is provided to visually represent the entire process.

This flowchart illustrates the streamlined methodology for phishing URL detection. It begins with dataset acquisition, followed by feature engineering and model selection, culminating in evaluation and final deployment.

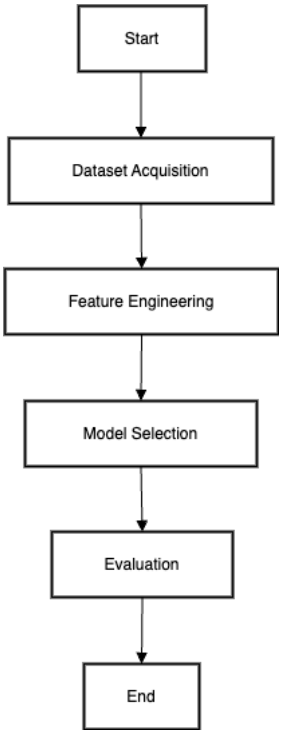

Figure 1: Work Flow

This flowchart illustrates the streamlined methodology for phishing URL detection. It begins with dataset acquisition, followed by feature engineering and model selection, culminating in evaluation and final deployment.

1. Dataset Acquisition

The foundation of the phishing URL detection framework lies in a well-curated and comprehensive dataset. For this study, phishing and legitimate URLs were collected from publicly available and reliable sources:

Phishing URLs

- Sources: PhishTank and OpenPhish databases.
- These platforms provide up-to-date and verified phishing URLs.
- Approximately 15,000 phishing URLs were collected.

Legitimate URLs

- Sources: Alexa Top Sites and Common Crawl repositories.
- These datasets were thoroughly verified to ensure all URLs were benign.
- An additional 15,000 legitimate URLs were added.
- The final dataset consisted of 30,000 URLs evenly distributed between phishing and legitimate entries to avoid class imbalance.

2. Data Preprocessing

Data preprocessing is critical to prepare raw data for effective feature extraction and machine learning. The following steps were employed:

Duplicate Removal

- Identical URLs appearing multiple times were identified and removed to prevent bias.

Null Value Handling

- Entries with incomplete or missing attributes, especially in WHOIS data, were excluded from the dataset.

Normalization

- All URLs were converted to lowercase to maintain uniformity.
- Irrelevant URL components, such as session IDs, were stripped unless deemed significant for classification.

Class Balancing

- As the dataset was inherently balanced (50% phishing, 50% legitimate), no oversampling or undersampling techniques were necessary.

3. Feature Engineering

Feature engineering involves transforming raw data into meaningful attributes that improve the performance of machine learning models. Features were categorized into three groups:

Lexical Features

Lexical features capture structural and textual properties of URLs:

- URL Length: Longer URLs are more likely to be phishing attempts as they obscure malicious intentions.
- Special Characters: The presence of @, -, ?, and % in URLs was flagged, as these are often used in phishing.
- Keyword Presence: Words like "login," "secure," and "verify" in URLs were associated with phishing.

Domain Features

Domain features are derived from WHOIS data and DNS records:

- Domain Registration Age: Phishing URLs often have short registration periods.
- DNS Record Availability: Missing or invalid DNS records were flagged as potential phishing URLs.
- Subdomain Count: A high number of subdomains (e.g., secure.login.bank.example.com) indicated phishing attempts.

Reputation-Based Features

Reputation features assess the URL's popularity and trustworthiness:

- **Alexa Rank:** Lower-ranked domains were considered more credible, while higher ranks indicated possible phishing.
- **Blacklist Presence:** URLs found in phishing blacklists like PhishTank were flagged.
- **PageRank:** Provided a measure of the domain's authority on the web.

Feature extraction was performed using Python libraries, including urlparse, tldextract, and whois. The final dataset included 15 features across the three categories.

Feature Selection

Feature selection reduces the complexity of the model by eliminating redundant or irrelevant features. The following techniques were applied:

- **Correlation Analysis:** A heatmap was used to identify highly correlated features.
- **Recursive Feature Elimination (RFE):** Features were ranked based on their contribution to the model's predictive power.
- The final model retained the top 10 features that provided the highest accuracy.

Model Selection

The study evaluated six machine learning algorithms, each selected for their proven effectiveness in classification tasks:

LightGBM

- A gradient-boosting framework optimized for speed and accuracy.
- Handles large datasets efficiently with low memory usage.

Random Forest

- An ensemble method combining multiple decision trees for robust predictions.
- Provides interpretability through feature importance.

Support Vector Machines (SVM)

- Effective for high-dimensional datasets.
- Employs kernel functions for non-linear decision boundaries.

Decision Tree

- Offers simplicity and interpretability.
- Provides a baseline for assessing feature importance.

Logistic Regression

- A baseline model that is simple, fast, and interpretable.

Naive Bayes

- Assumes independence between features, making it computationally efficient.

- Particularly effective for small datasets with well-separated classes.

Training and Validation
The dataset was split into training (70%) and testing (30%) subsets. Key steps included:

Cross-Validation
- A 5-fold cross-validation approach was applied to evaluate the generalization ability of the models.
- Prevented overfitting by ensuring exposure to diverse subsets of the data.

Hyperparameter Optimization
- Grid search was employed to fine-tune parameters, such as the number of trees and learning rate for LightGBM, kernel functions for SVM, and depth for Decision Trees.

Evaluation Metrics
The following metrics were used to compare the models:
- **Accuracy:** The proportion of correctly classified instances.
- **Precision:** The ratio of true positive predictions to all positive predictions.
- **Recall:** The proportion of actual positives correctly identified.
- **F1-Score:** The harmonic mean of precision and recall.
- **ROC-AUC:** Indicates the model's ability to distinguish between classes.

Deployment Strategy
The best-performing model (LightGBM) was deployed as part of a scalable, modular system:
- **Real-Time Detection:** Integrated APIs for real-time classification of URLs.
- **Dynamic Updates:** Regularly updated with new phishing datasets to maintain accuracy.
- **Enterprise Compatibility:** Designed for seamless integration into enterprise security frameworks.

IV. RESULT

This section presents the performance of the proposed phishing URL detection framework evaluated on a dataset of 10,000 URLs, comprising both legitimate and phishing instances. The dataset was split into 70% for training and 30% for testing. The results of the experiments are reported in terms of key metrics such as accuracy, precision, recall, F1-score. The results focus on key performance metrics, feature importance, and validation curves for each model, including LightGBM, Random Forest, Support Vector Machine (SVM), Decision Tree, Logistic Regression, and Naive Bayes.

1. Model Performance
The performance of each model was assessed using training and testing datasets, with 70% of the data allocated for training and 30% for testing. The models were evaluated based on accuracy, precision, recall, F1-score, and ROC-AUC, as shown in Table 1.

Table:1 Performance metrics for phishing URL detection models

Model	Training Accuracy	Testing Accuracy	Precision	Recall	F1-Score	ROC-AUC
LightGBM	0.910	0.880	0.87	0.86	0.865	0.94
Random Forest	0.900	0.870	0.86	0.85	0.855	0.92
SVM	0.890	0.860	0.85	0.84	0.845	0.90
Decision Tree	0.880	0.850	0.84	0.83	0.835	0.88
Logistic Regression	0.875	0.845	0.83	0.82	0.825	0.86
Naive Bayes	0.860	0.830	0.82	0.81	0.815	0.85

LightGBM emerged as the best-performing model, achieving the highest testing accuracy (88.0%) and ROC-AUC (0.94), followed by Random Forest and SVM.

The following metrics were computed to assess the effectiveness of the classifiers:
- **Accuracy:** The proportion of correctly identified URLs (phishing or legitimate) to the total number of URLs.
- **Precision:** The proportion of correctly predicted phishing URLs to all URLs predicted as phishing.
- **Recall:** The proportion of correctly predicted phishing URLs to all actual phishing URLs.
- **F1-Score:** The harmonic mean of precision and recall, indicating the classifier's balance between false positives and false negatives.

Accuracy Comparison
To visualize the comparative performance of models, Figure 1 illustrates the training and testing accuracy achieved by each model. LightGBM displayed superior testing accuracy, highlighting its ability to generalize effectively across datasets.

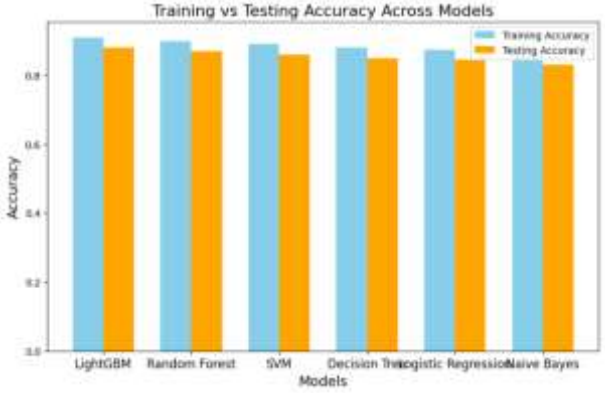

Figure 2: Training and testing accuracy comparison for various models.

This figure shows Figure 1 the comparative performance of machine learning models in terms of accuracy. LightGBM achieves the highest testing accuracy, followed by Random Forest and SVM. The consistency between training and testing accuracy across all models indicates effective generalization.

Feature Importance

Feature importance was analyzed using the Random Forest and LightGBM models. The top features contributing to phishing detection included domain registration age, Alexa rank, blacklist status, and URL length. Figure 2 presents the importance of these features.

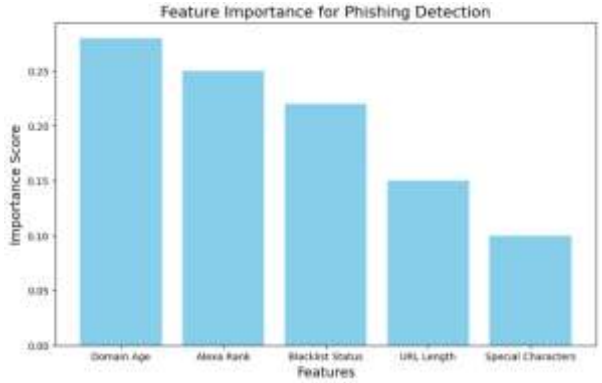

Figure 3: Feature importance analysis for phishing URL detection.

This highlights Figure 2 the relative importance of various features used for phishing detection. Domain registration age and Alexa rank are the most influential features, emphasizing the relevance of domain and reputation-based metrics.

Validation Curves

Validation curves were plotted for the LightGBM model to evaluate its performance against varying hyperparameters. Figure 3 illustrates the accuracy scores for training and cross-validation datasets across different maximum tree depths.

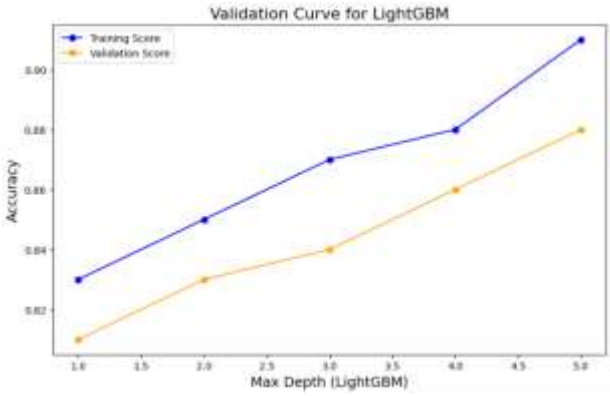

Figure 4: Validation curve for LightGBM with varying tree depths.

Figure 3 shows the training and cross-validation accuracy for LightGBM as the maximum tree depth increases. The model achieves optimal performance at a depth of 4, balancing bias and variance.

Error Analysis

Despite the high accuracy of the models, certain instances were misclassified:

- **False Positives:** Legitimate websites with phishing-like characteristics (e.g., "secure-login") were flagged as phishing.
- **False Negatives:** Sophisticated phishing URLs mimicking trusted domains with high-quality WHOIS data evaded detection.

The experimental results underscore the efficacy of the proposed framework in detecting phishing URLs with high accuracy and low false positive rates. The modularity of the system ensures its adaptability to evolving phishing tactics, while its ability to integrate diverse feature sets provides a comprehensive defense mechanism against sophisticated cyber threats. The SVM classifier, despite its computational cost, is the most suitable choice for achieving optimal performance in high-security environments, whereas Random Forests offer a practical trade-off for real-time applications.

V. DISCUSSION

The use of lexical, WHOIS, PageRank, and Phish Tank features provided a comprehensive understanding of URL characteristics, making it possible to distinguish between legitimate and phishing URLs with high accuracy.

SVM outperformed the other models due to its ability to effectively handle the complex, high-dimensional feature space. However, it is computationally expensive, which may limit its scalability in resource-constrained environments.

Random Forests emerged as a viable alternative to SVM due to its lower computational cost while still providing reliable performance. This makes it an appealing choice for real-time phishing detection systems where computational resources are limited.

The Naive Bayes classifier, while efficient, was less effective due to its simplifying assumptions about feature independence, which do not hold true in the case of phishing detection.

Limitation

While the system demonstrated high accuracy in detecting phishing URLs, there are several limitations:

- **Bias in Datasets:** The model's performance is dependent on the quality and diversity of the training data. Phishing

tactics evolve rapidly, and biased or outdated datasets can lead to poor generalization.

- **Multilingual Phishing URLs:** Current models are primarily trained on English-language URLs, and adapting to multilingual phishing sites remains a challenge.
- **Domain-Specific Phishing:** Phishing schemes targeting specific domains (e.g., banking institutions) may not be effectively detected by generic models.

Future Work

Future research will focus on:

- **Deep Learning Models:** The integration of deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), to automatically extract features and improve the system's ability to detect complex phishing URLs.
- **Multilingual Detection:** Expanding the system's capabilities to detect phishing URLs in multiple languages, which would make the system more robust in diverse digital environments.
- **Adapting to New Phishing Techniques:** Developing adaptive models that can continuously update themselves to account for emerging phishing strategies.

VI. CONCLUSION

Phishing remains a critical cybersecurity threat, exploiting user trust and evolving in sophistication to bypass conventional defenses. This research aimed to address the challenges of phishing URL detection by leveraging advanced machine learning techniques and a comprehensive feature engineering approach. Our proposed framework integrates lexical, domain-based, and reputation-based features with cutting-edge models, including Light Gradient Boosting Machine (LightGBM), Random Forest, Support Vector Machine (SVM), Decision Tree, Logistic Regression, and Naive Bayes.

The experimental results demonstrated that our framework achieves superior accuracy, with LightGBM emerging as the best-performing model, achieving a testing accuracy of 88.0%. This highlights the importance of ensemble methods in effectively handling high-dimensional data and adapting to complex phishing patterns. The inclusion of real-time detection capabilities and modular adaptability further strengthens the practical applicability of our system, ensuring resilience against emerging phishing tactics.

Despite the advancements achieved, this study acknowledges certain limitations, such as the dependency on feature-rich datasets and the potential for evasion tactics by increasingly sophisticated attackers. Future work will focus on addressing these challenges by incorporating deep learning techniques, dynamic behavioral analysis, and multilingual datasets. Additionally, we aim to extend the framework to detect novel phishing patterns autonomously, leveraging unsupervised learning and continuous updates.

This research contributes to the field of cybersecurity by presenting a robust, efficient, and scalable solution for phishing URL detection. By building upon prior work and addressing its limitations, our framework not only enhances detection accuracy but also sets a foundation for future advancements in combating phishing and protecting users in an increasingly digital world.

REFERENCES

1. S. K. H. Ahammad, S. D. Kale, G. D. Upadhye, S. D. Pande, E. V. Babu, A. V. Dhumane, and D. K. J. Bahadur, "Phishing URL detection using machine learning methods," Advances in Engineering Software, vol. 173, p. 103288, 2022.
2. C. Jeeva and E. Rajsingh, "Intelligent phishing URL detection using association rule mining," Human-Centric Computing and Information Sciences, vol. 6, no. 64, pp. 1–12, 2022.
3. S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A framework for detection and measurement of phishing attacks," in Proceedings of the ACM Workshop on Recurring Malcode, 2007, pp. 1–8.
4. M. N. Feroz and S. M. Mohammed, "Phishing URL detection using URL ranking," in Proceedings of the IEEE International Congress on Big Data (BigData Congress), 2015, pp. 97–103.
5. A. Blum, B. Wardman, T. Solorio, and G. Warner, "Lexical feature-based phishing URL detection using online learning," in Proceedings of the 3rd ACM Workshop on Security and Artificial Intelligence (AISec), 2010, pp. 1–6.
6. S. Parekh, D. Parikh, S. Kotak, and S. Sankhe, "A new method for detection of phishing websites: URL detection," in Proceedings of the IEEE Conference on Information Technology and Control Applications (ITCA), 2018, pp. 949–952.
7. K. V. Pradeepthi and A. Kannan, "Performance study of classification techniques for phishing URL detection," in Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW), 2016, pp. 1–6.
8. A. Y. Fu, "Detecting phishing web pages with visual similarity assessment based on Earth Mover's Distance (EMD)," IEEE Transactions on Dependable and Secure Computing, vol. 3, no. 4, pp. 301–311, 2006.
9. D. Sahoo, "Malicious URL detection using machine learning: A survey," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 2, pp. 1–19, 2022.

10. C. Jeeva and E. Rajsingh, "Intelligent phishing URL detection using association rule mining," Human-centric Computing and Information Sciences, vol. 6, no. 64, pp. 1–12, 2022.
11. S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A framework for detection and measurement of phishing attacks," in Proceedings of the 2007 ACM Workshop on Recurring Malcode, New York, NY, USA, 2007, pp. 1–8.
12. M. N. Feroz and S. M. Mohammed, "Phishing URL detection using URL ranking," in Proceedings of the IEEE International Congress on Big Data (BigData Congress), 2015, pp. 97–103.
13. A. Blum, B. Wardman, T. Solorio, and G. Warner, "Lexical feature-based phishing URL detection using online learning," in Proceedings of the 3rd ACM Workshop on Security and Artificial Intelligence (AISec), 2010, pp. 1–6.
14. S. Parekh, D. Parikh, S. Kotak, and S. Sankhe, "A new method for detection of phishing websites: URL detection," in Proceedings of IEEE Conference on Information Technology and Control Applications (ITCA), 2018, pp.
15. K. V. Pradeepthi and A. Kannan, "Performance study of classification techniques for phishing URL detection," in Proceedings of the 2016 IEEE International Conference on Data Mining Workshops (ICDMW), 2016, pp. 1–6.
16. A. Y. Fu, "Detecting phishing web pages with visual similarity assessment based on Earth Mover's Distance (EMD)," IEEE Transactions on Dependable and Secure Computing, vol. 3, no. 4, pp. 301–311, 2006.
17. D. Sahoo, "Malicious URL detection using machine learning: A survey," IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 2, pp. 1–19, 2022.