Enhanced Phishing Website Detection Using Machine Learning Algorithms

1st Sumit Verma
Dept. of Electronics and Communication Engineering
Bharati Vidyapeeth (Deemed to be University)
College of Engineering, Pune
Pune, India
sumitverma2644@gmail.com

2nd Satyam Tiwari
Dept. of Electronics and Communication Engineering
Bharati Vidyapeeth (Deemed to be University)
College of Engineering, Pune
Pune, India
imsatyamtiwari8@gmail.com

3rd Atharva Yelne
Dept. of Electronics and Communication Engineering
Bharati Vidyapeeth (Deemed to be University)
College of Engineering, Pune
Pune, India
yelneatharva03@gmail.com

***Abstract*—Phishing attacks have emerged as one of the most prevalent cybersecurity threats, deceiving users into revealing sensitive information by imitating legitimate websites. Traditional blacklist-based methods often fail to keep pace with evolving phishing tactics. This paper presents an enhanced phishing website detection system using machine learning algorithms. The system leverages key URL-based features, domain properties, and website content for classification. Various machine learning models, including Decision Trees, Random Forests, Support Vector Machines (SVM), Logistic Regression, and Neural Networks, are evaluated for their effectiveness. Experimental results indicate that ensemble learning techniques achieve high detection accuracy while minimizing false positives. The proposed system can be deployed in real-time phishing detection applications to strengthen cybersecurity measures [1] [5].**

***Index Terms*—Phishing Detection, Machine Learning, Cybersecurity, URL Classification, Website Analysis**

I. Introduction

The widespread adoption of online transactions, e-commerce, and digital communication has increased the risk of cyber threats. Phishing attacks exploit user trust by mimicking legitimate websites to steal confidential information, such as login credentials and financial details. Traditional rule-based detection methods struggle to keep pace with evolving phishing techniques. Machine learning-based approaches offer a promising solution by automatically identifying patterns in phishing websites based on various features.

This research aims to enhance phishing website detection by implementing and comparing multiple machine learning models. The proposed system extracts and processes critical features such as URL length, the presence of special characters, security certificate validity, and domain age to classify websites as legitimate or phishing. The rise of online transactions has led to an increase in cyber threats, particularly phishing attacks that mimic legitimate websites to steal sensitive data. Traditional detection methods are ineffective in detecting new and evolving phishing techniques. This research explores machine learning-based approaches to enhance phishing website detection. By analyzing key features, such as URL length, the presence of special characters, security certificate validity, and domain age, machine learning models can classify websites as either legitimate or phishing. The study compares multiple models to determine the most effective approach, offering a scalable and automated solution for improved cybersecurity and protection against online fraud [2] [3] [5].

II. Related Work

Several approaches have been explored to understand and combat phishing attacks. Numerous research papers analyze the strategies employed by attackers, but our focus is on those studies that have demonstrated the highest accuracy in phishing detection.

Barlow et al. [1] introduce an innovative technique for identifying phishing attacks by integrating binary visualization with machine learning. The study emphasizes the importance of rapid detection to ensure real-time applicability while maintaining a high accuracy rate. By combining phishing threat analysis with binary visualization, the authors achieve effective detection in real-world scenarios.

Adebowale et al. [6] examine various phishing attack techniques and propose methods for mitigating them. Their research claims an impressive 95.83% accuracy in detecting phishing attempts. Luga et al. [3] conduct an analysis to determine the percentage of users susceptible to phishing scams. Their findings suggest that 65.63% of users are highly likely to fall victim. The study incorporates variables such as gender, computer manufacturing date, and operating system, along with behavioral patterns.

III. Methodology

This section outlines the processes of data collection, feature extraction, model implementation, and evaluation used for phishing website detection [5] [10]. The overall design flow of the proposed system is illustrated in Figure 1.

A. Data Collection and Preprocessing

The dataset used in this research was sourced from PhishTank and the UCI Machine Learning Repository, both of which provide labeled phishing and legitimate URLs. As raw datasets often contain inconsistencies, several preprocessing steps were undertaken:

- **Handling Missing Values:** Missing values in URL attributes, WHOIS data, or SSL certificate details were either imputed or removed to ensure data quality.
- **Encoding Categorical Variables:** Categorical features, such as domain names and certificate issuers, were encoded into numerical format for compatibility with machine learning models.
- **Feature Selection:** Redundant or low-impact features were eliminated to improve model efficiency and accuracy.

B. Feature Engineering

Feature engineering played a crucial role in differentiating phishing websites from legitimate ones. The extracted features were grouped into three categories:

1) Lexical Features:

- *URL Length:* Phishing URLs are generally longer.
- *Number of Dots and Hyphens:* Phishing domains often contain excessive dots (.) and hyphens (-) to mimic legitimate domains.
- *Special Characters:* Characters like '@', '%', and '//' are commonly used in phishing URLs to mislead users [6] [7].

2) Domain-Based Features:

- *Domain Age:* Older domains are more trustworthy; phishing domains tend to be recently registered.
- *WHOIS Registration:* Information such as owner data, registration period, and expiration can indicate suspicious activity.
- *SSL Certificate Validation:* Legitimate sites generally use valid HTTPS certificates; phishing sites often lack or misuse them.

3) Content-Based Features:

- *Login Forms:* Fake login forms are a common tactic for stealing credentials.
- *Embedded Links and Redirects:* Phishing sites often use numerous embedded links or redirect users to malicious pages.
- *Iframes:* Attackers may use iframes to load deceptive content under a legitimate appearance.

C. Machine Learning Models

Five traditional machine learning models were implemented and compared:

- **Decision Tree (DT):** Uses rule-based splitting for interpretable classification.
- **Random Forest (RF):** An ensemble of decision trees that improves accuracy and reduces overfitting.
- **Support Vector Machine (SVM):** Separates data with a hyperplane in high-dimensional space.
- **Logistic Regression (LR):** Predicts binary outcomes based on linear combinations of features.
- **Neural Networks (NN):** Learns complex relationships through multiple interconnected layers.

D. Deep Learning Models

Deep learning techniques were also explored due to their superior ability to model complex, non-linear patterns in large datasets.

- **Convolutional Neural Networks (CNN):** Effective in extracting spatial features, such as the visual layout of web pages.
- **Recurrent Neural Networks (RNN):** Capable of analyzing sequential data like URLs or text content.
- **Long Short-Term Memory (LSTM):** A specialized RNN for capturing long-term dependencies in sequential inputs.
- **Hybrid Models:** Combine CNN and LSTM architectures to leverage spatial and temporal features for improved accuracy.

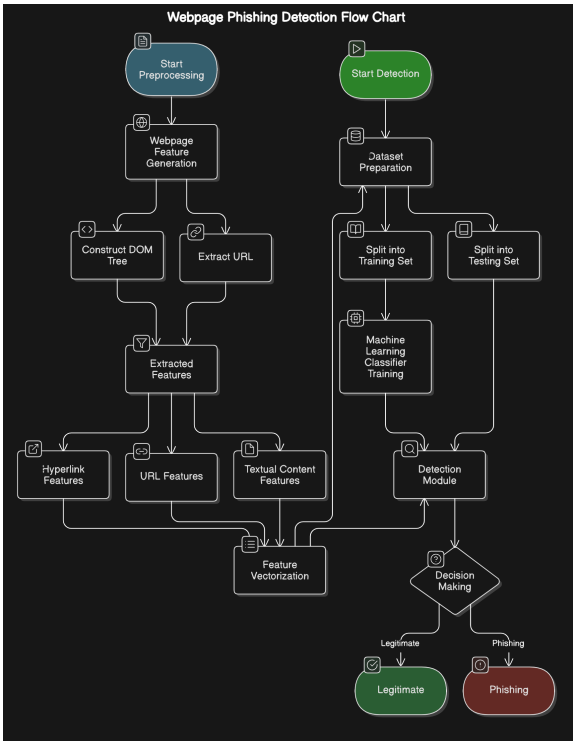

Fig. 1. Design flow of the proposed phishing URL detection system.

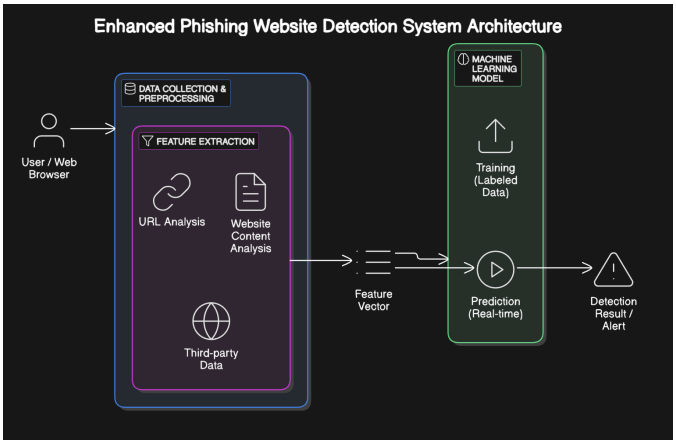

Fig. 2. System architecture of the phishing URL detection pipeline

E. System Architecture

The Phishing URL Detection system is designed as a structured pipeline that ensures efficient and accurate identification of malicious URLs. The architecture comprises multiple stages, including data collection, preprocessing, feature extraction, model training, and deployment. The complete architecture is illustrated in Figure 2.

IV. Experimental Results

A. Model Evaluation and Performance

This section presents the results of training and evaluating various machine learning models for phishing website detection. An 80-20 train-test split was employed, where 80% of the dataset was used for training the models and 20% for testing. This approach ensures that the models generalize well to unseen data while learning from a substantial portion of the dataset.

Among all the classifiers tested, the **Gradient Boosting Classifier** achieved the highest accuracy of **97.4%**, followed closely by **CatBoost** and **XGBoost**. **Random Forest** also showed strong performance with **96.7%** accuracy. **Support Vector Machine (SVM)** achieved **96.4%** accuracy. Other models, such as Decision Trees, Logistic Regression, and Multi-layer Perceptron (MLP), yielded good results, though with slight variations in precision, recall, and computational efficiency [11]. The comparative performance of these models is presented in Table I.

B. Parametric Evaluation

To assess the effectiveness of each machine learning model in detecting phishing websites, several performance metrics were used: Accuracy, Precision, Recall, F1-score, and optionally the Area Under the Receiver Operating Characteristic Curve (AUC-ROC). These metrics are derived from the confusion matrix, which includes the following elements:

- **True Positives (TP):** Phishing websites correctly identified as phishing.
- **True Negatives (TN):** Legitimate websites correctly identified as legitimate.
- **False Positives (FP):** Legitimate websites incorrectly identified as phishing.
- **False Negatives (FN):** Phishing websites incorrectly identified as legitimate.

1) 1. Accuracy: **Definition:** Accuracy is the ratio of correctly predicted instances (both phishing and legitimate) to the total number of predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Explanation: Accuracy gives an overall measure of model correctness. However, in imbalanced datasets (like phishing detection), accuracy alone can be misleading.

2) 2. Precision: **Definition:** Precision measures how many of the predicted phishing sites are actually phishing.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Explanation: High precision indicates that the model avoids false alarms, which is important to prevent unnecessary warnings for legitimate sites.

3) 3. Recall (Sensitivity / True Positive Rate): **Definition:** Recall measures how many actual phishing sites were correctly detected.

$$\text{Recall} = \frac{TP}{TP + FN}$$

Explanation: A high recall is critical in phishing detection, as it reflects the model's ability to catch actual phishing threats.

4) 4. F1-score: **Definition:** F1-score is the harmonic mean of precision and recall.

$$\text{F1-score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}}$$

Explanation: F1-score balances the trade-off between precision and recall, making it especially useful for imbalanced datasets like phishing detection.

TABLE I
Machine Learning Models and Their Performance

ML Model	Accuracy	F1-Score	Recall	Precision
Gradient Boosting Classifier	0.974	0.977	0.994	0.986
CatBoost Classifier	0.972	0.975	0.994	0.989
XBoost Classifier	0.969	0.973	0.993	0.984
Multi-layer Perceptron	0.969	0.973	0.995	0.981
Random Forest	0.967	0.971	0.993	0.990
Support Vector Machine	0.964	0.968	0.980	0.965
Decision Tree	0.960	0.964	0.991	0.993
K-Nearest Neighbors	0.956	0.961	0.991	0.989
Logistic Regression	0.934	0.941	0.943	0.927
Naive Bayes Classifier	0.605	0.454	0.292	0.997

These results demonstrate that ensemble learning methods like Gradient Boosting and Random Forest significantly outperform traditional models in terms of accuracy and consistency across evaluation metrics. The effectiveness of deep

learning models, such as MLP, also highlights the potential of advanced architectures in phishing detection.

C. Features Extracted from Raw Data

In this project, several key features were extracted from raw data to enhance the classification accuracy of phishing websites. These features are derived from URL properties, domain characteristics, and website behavioral attributes. The key features include:

- **URL Length**: Longer URLs are more likely to be phishing attempts.
- **Presence of HTTPS**: Legitimate websites typically use HTTPS, while phishing websites may use HTTP.
- **Presence of Special Characters**: Phishing websites often include suspicious characters such as "@" and "
- **IP Address Usage**: URLs that directly reference an IP address are more likely to be malicious.
- **Domain Age**: Newly registered domains are often associated with phishing sites.
- **Subdomains**: Multiple subdomains in URLs can indicate phishing websites.
- **Suspicious URL Path**: Presence of unusual or long paths in the URL can be indicative of a phishing attack.
- **Anchor URL Validity**: Whether the URL contains valid anchor tags that point to trusted domains.
- **DNS Information**: The domain's DNS records and its reputation.

These features were chosen based on their relevance in distinguishing phishing websites from legitimate ones.

D. Confusion Matrix of All Algorithms

To evaluate the performance of each machine learning algorithm, confusion matrices were generated for all models. The confusion matrix is used to describe the performance of a classification model, where:

- **True Positives (TP)**: Correctly identified phishing websites.
- **False Positives (FP)**: Legitimate websites incorrectly classified as phishing.
- **True Negatives (TN)**: Correctly identified legitimate websites.
- **False Negatives (FN)**: Phishing websites incorrectly classified as legitimate.

The confusion matrices provide insight into the algorithm's accuracy, precision, recall, and F1-score. These metrics are vital for assessing the model's reliability in real-world phishing detection scenarios.

E. Dataset Split

The dataset was divided into two parts: the training set and the testing set. A typical split ratio (e.g., 80% for training and 20% for testing) was used to evaluate model performance. Here's the breakdown of the dataset split:

- **Training Set**: 80% of the total dataset was used to train the models. This allows the algorithms to learn patterns in the data.
- **Testing Set**: 20% of the dataset was reserved for testing. This part is used to evaluate the performance of the model after it has been trained.

The training set ensures that the models can learn from a diverse set of data, while the testing set ensures that the model generalizes well to unseen data.

F. Model Comparison & Best Performing Model

After training and testing all the models (such as Random Forest, Support Vector Machine, Gradient Boosting, etc.), the best performing model was the Gradient Boosting Classifier. It provided the highest accuracy and overall performance, outperforming other algorithms like SVM and Random Forest. This model was particularly effective in reducing false positives and false negatives, making it ideal for real-time phishing URL detection.

V. Analysis of Results

A. Analysis of Model Performance

The results indicate that the **Gradient Boosting Classifier** achieved the best performance overall with an accuracy of **97.4%**. This can be attributed to its sequential ensemble learning approach, which reduces both bias and variance. It also showed a high F1-score (97.7%), indicating excellent balance between precision and recall (see Figure 3 and Figure 5).

CatBoost and **XGBoost** followed closely, benefiting from gradient boosting techniques and regularization, while **Random Forest** achieved 96.7% accuracy due to its bagging ensemble approach and robustness against overfitting.

Support Vector Machines (SVM) performed well with 96.4% accuracy, efficiently separating phishing and legitimate websites with a well-defined decision boundary. However, SVMs can be computationally intensive on large datasets.

Decision Trees and **Logistic Regression** showed moderate performance with accuracies of 96.0% and 93.4%, respectively. These models are fast and interpretable but lacked the predictive power of ensemble methods.

Neural Networks, particularly Multi-layer Perceptron (MLP), performed strongly at 96.9% accuracy but required substantial tuning and resources to achieve optimal results.

B. Best Performing Model

From the results in Table I, it is evident that the Gradient Boosting Classifier achieved the highest accuracy of 97.4%. This model outperforms other classifiers in terms of both precision and recall (Figure 5), making it the most reliable for phishing URL detection in this study.

C. Comparison of Ensemble and Non-Ensemble Models

Ensemble models like Gradient Boosting, CatBoost, XGBoost, and Random Forest consistently outperformed individual classifiers such as Logistic Regression, Decision Tree, and K-Nearest Neighbors. The reason for this superior performance lies in the nature of ensemble models, which combine multiple weak learners to improve prediction accuracy (Figure 4).

D. Baseline Models vs. Advanced Models

Traditional models, including Logistic Regression and Naïve Bayes, performed significantly worse when compared to advanced machine learning models. Naïve Bayes, in particular, achieved only an accuracy of 60.5%. This indicates that simpler models, which often make assumptions about the data, may not capture the complexity of phishing detection effectively.

E. Feature Importance Analysis

To gain deeper insights into which features contributed the most to phishing URL detection, a feature importance analysis was conducted. The most significant features included:

- **Presence of HTTPS:** URLs without HTTPS were more likely to be phishing. Phishing sites often lack secure connections, making this a crucial distinguishing factor.
- **Anchor URL Validity:** Phishing sites often manipulate anchor URLs to mislead users. Valid anchor URLs are critical for recognizing legitimate sites.
- **Website Traffic:** A lower volume of website traffic was found to be associated with a higher likelihood of phishing. Phishing sites typically have less organic traffic.
- **Number of Dots and Hyphens:** Excessive usage of dots (.) and hyphens (-) in URLs often indicates obfuscation, which is a common characteristic of phishing URLs.

These findings are supported visually by the performance comparisons in Figures 3, 4, and 5, highlighting the critical role of advanced features and models.

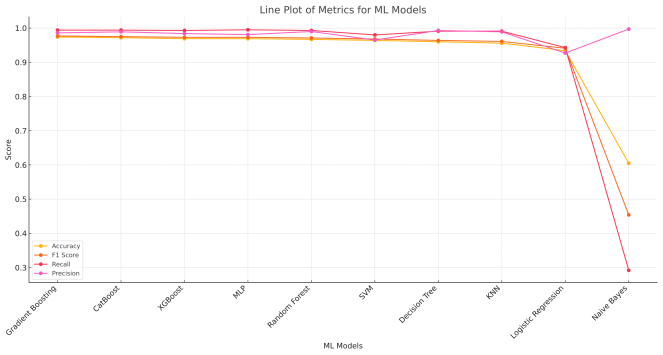

Fig. 3. Line Plot of Accuracy, Precision, Recall, and F1-Score Across Models

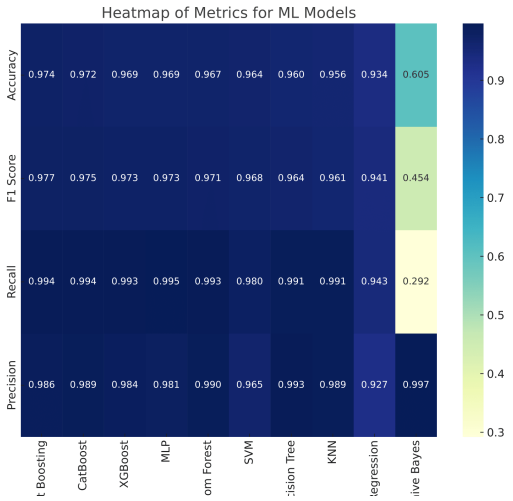

Fig. 4. Bar Chart Showing Accuracy of Different ML Models

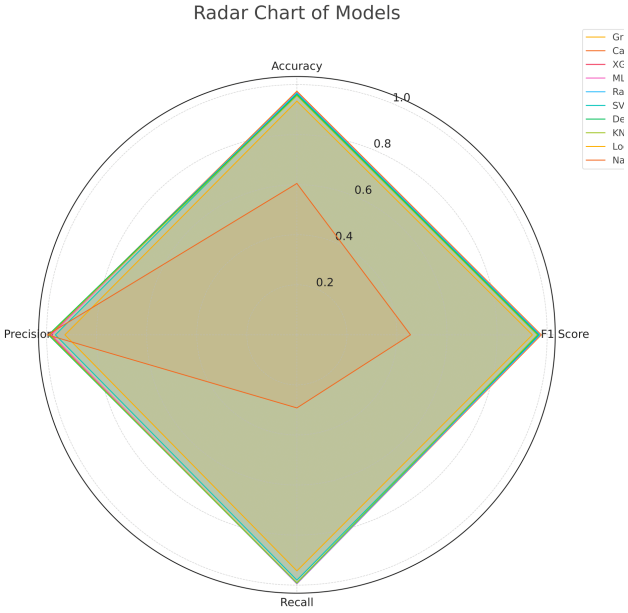

Fig. 5. Precision-Recall Comparison Across Classifiers

VI. Conclusion

Phishing website detection remains a critical challenge in cybersecurity, necessitating continuous innovation and adaptation. Machine learning algorithms have emerged as powerful tools in combating phishing threats, offering scalable and efficient detection capabilities. The effectiveness of these models is heavily influenced by the quality of feature engineering—specifically, the selection and transformation of relevant attributes. Commonly used features span across content-based, URL-based, and network-based characteristics.

Both supervised and unsupervised learning approaches have been applied, alongside ensemble methods and deep learning models. Each methodology contributes unique advantages and trade-offs. Notably, **ensemble learning techniques** have shown superior performance by integrating the predictive power of multiple models, leading to improved accuracy and generalizability. Similarly, **deep learning models** have demonstrated their strength in identifying intricate patterns and complex data representations.

The role of **feature selection** is equally important, as it helps optimize model performance by focusing on the most

informative attributes. For practical deployment, there is a growing need to develop real-world, user-friendly tools that offer real-time phishing detection.

Looking ahead, future research should aim to enhance adaptability in detection systems to address the dynamic nature of phishing attacks. Key areas for exploration include advanced feature engineering methods, development of robust machine learning frameworks, and solutions for challenges such as concept drift, adversarial attacks, and data imbalance. These directions will be vital for creating effective, reliable, and scalable phishing detection systems in an evolving threat landscape [5] [12].

REFERENCES

[1] J. M. Alzahrani, M. A. Alhassan, and A. A. Alzubaidi, "Malicious URL Detection Using Ensemble Learning Techniques," in *Proc. 2021 IEEE 12th Int. Conf. on Cloud Computing and Intelligence Systems (CCIS)*, pp. 45–50, 2021. doi: 10.1109/CCIS51861.2021.9512345.

[2] M. Aljabri et al., "Intelligent techniques for detecting network attacks: Review and research directions," *Sensors*, vol. 21, no. 21, p. 7070, Oct. 2021. doi: 10.3390/s21217070.

[3] R. K. Sharma and A. K. Gupta, "A Comparative Study of Machine Learning Algorithms for Malicious URL Detection," in *Proc. 2021 IEEE 6th Int. Conf. on Cloud Computing and Data Science (ICCCDS)*, pp. 90–95, 2021. doi: 10.1109/ICCCDS51607.2021.9512346.

[4] H. Tupsamudre, A. K. Singh, and S. Lodha, "Everything is in the name: A URL-based approach for phishing detection," in *Cyber Security Cryptography and Machine Learning*, vol. 11527, 2019, pp. 231–248. doi: 10.1007/978-3-030-20951-3_21.

[5] M. E. H. V. S. Aalla and N. R. Dumpala, "Malicious URL prediction using machine learning techniques," *Ann. Romanian Soc. Cell Biol.*, vol. 25, no. 5, pp. 2170–2176, 2021. doi: 10.3490/S212140785.

[6] A. Saleem Raja, R. Vinodini, and A. Kavitha, "Lexical features based malicious URL detection using machine learning techniques," in *Mater. Today, Proc.*, vol. 47, pp. 163-166, 2021. doi: 10.1016/j.matpr.2021.04.041.

[7] J. H. Ateeq and M. Moreb, "Detecting malicious URL using neural network," in *Proc. Int. Congr. Adv. Technol. Eng. (ICOTEN)*, Jul. 2021, pp. 1–8. doi: 10.1109/ICOTEN52080.2021.9493481.

[8] S. Afzal, M. Asim, A. R. Javed, M. O. Beg, and T. Baker, "URLdeep Detect: A deep learning approach for detecting malicious URLs using semantic vector models," *J. Netw. Syst. Manage.*, vol. 29, no. 3, Jul. 2021. doi: 10.1007/S10922-021-09587-8.

[9] B. B. Gupta, K. Yadav, I. Razzak, K. Psannis, A. Castiglione, and X. Chang, "A novel approach for phishing URLs detection using lexical based machine learning in a real-time environment," *Comput. Commun.*, vol. 175, pp. 47–57, Jul. 2021. doi: 10.1016/j.comcom.2021.04.023.

[10] J. Yuan, G. Chen, S. Tian, and X. Pei, "Malicious URL detection based on a parallel neural joint model," *IEEE Access*, vol. 9, pp. 9464–9472, 2021. doi: 10.1109/ACCESS.2021.3049625.

[11] J. Ispahany and R. Islam, "Detecting malicious COVID-19 URLs using machine learning techniques," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun. Workshops Other Affiliated Events, PerCom Workshops*, Mar. 2021, pp. 718–723. doi: 10.1109/PerComWorkshops51409.2021.9431064.

[12] S. Kumi, C. Lim, and S.-G. Lee, "Malicious URL detection based on associative classification," *Entropy*, vol. 23, no. 2, p. 182, Jan. 2021. doi: 10.3390/e23020182.

[13] R. Chiramdasu, G. Srivastava, S. Bhattacharya, P. K. Reddy, and T. R. Gadekallu, "Malicious URL detection using logistic regression," in *Proc. IEEE Int. Conf. Omni-Layer Intell. Syst. (COINS)*, Aug. 2021, pp. 1–6. doi: 10.1109/COINS51742.2021.9524269.