International Journal of Computer Science Engineering Techniques – Volume 9
Issue 6, November - December - 2025

Safeguarding Social Media Users through Machine Learning–Powered Fake URL Detection Systems

[1]Mohd Athar Nawab Jani, [2]A.Sumana, [3]B.Vinuthna, [4]B.Srinithi, [5]A.Keerthana, [6]B.Nishwitha
[1]Assistant Professor, [2,3,4,5,6]UG Student, [1,2,3,4,5,6]Department of CSE (Data Science) Malla Reddy Engineering Collge for Women, Secunderabad, India
E-Mail: sumanaangajala282@gmail.com

Abstract
Counterfeit web addresses have become a major instrument of disseminating false information on social media. The importance of identifying these counterfeit URLs is to ensure that the propagation of fake information is prevented as well as the reliability of social media networks. In our paper, we suggest a machine literacy-based solution to identify fake URLs in the social media. We trained and approximate several machine learning models such as decision trees, arbitrary timbers and support vector machines on the dataset of real and fake URLs gathered on social media websites. We have found that the arbitrary timber algorithm had the highest delicacy of 96.5 compared to the other algorithms. We also examined the usefulness of colour features comparable to the length of URLs, sphere name, and URL order, and implemented the sphere name and URL order to be the most didactic features in identifying fake URLs. Our solution offers a reliable and efficient outcome to ascertain counterfeit URLs in the social media that can be applied to curb the diffusion of fake data and the reliability of social media networks.

Keywords: Fake URL, Machine learning, Phishing, Spamming, Malware, Lexical Features, Cat Boost classifier, Gradient boosting classifier, SPM, Decision tree, Logistic Regression, Naive Bayes classifier, KNN.

1 INTRODUCTION

In the digital era, social media platforms have become powerful tools for communication, marketing, and information exchange. However, their open and dynamic nature has also made them fertile ground for cyber threats, including phishing, misinformation campaigns, and the spread of malicious or fake URLs. These deceptive links often lead unsuspecting users to fraudulent websites designed to steal credentials, distribute malware, or manipulate public opinion. According to recent cybersecurity reports, over 40% of phishing attempts now originate from social media channels, posing a severe threat to user privacy and digital trust.

Traditional URL blacklisting and rule-based filtering methods struggle to keep pace with the growing sophistication of fake links. Attackers frequently modify URLs using obfuscation techniques such as domain spoofing, shortening services, or hidden redirects, which render static defense systems ineffective. Consequently, there is an urgent need for adaptive and intelligent mechanisms capable of learning from evolving data patterns to detect and block harmful URLs in real time.

Machine learning (ML) has emerged as a promising approach to address this challenge. By analyzing structural, lexical, and behavioral features of URLs—such as domain age, keyword composition, redirection patterns, and content attributes—ML models can automatically classify links as legitimate or malicious. Various algorithms including Support Vector Machines (SVM), Random Forests, Logistic Regression, and Deep Neural Networks have demonstrated remarkable accuracy in identifying phishing and fake URLs compared to traditional heuristic-based systems. Integrating these models into social media ecosystems enables proactive detection, reducing user exposure to online fraud.

The emergence of new communication technologies has greatly impacted the growth and promotion of

businesses across various applications. However, with these advancements comes the use of sophisticated techniques to attack and scam users. Cyber-attacks such as malicious websites that sell counterfeit goods, steal sensitive information, and install malware have become increasingly common. These attacks can be carried out using a wide range of techniques, including explicit hacking attempts, phishing, man-in-the middle attacks, SQL injections, and more. Malicious URLs are often used to spread compromised content and are responsible for a significant portion of cyber-attacks. It is estimated that one-third of all websites are malicious in nature. A URL, which is the global address of all documents and resources on the World Wide Web, comprises a protocol identifier and a resource name that specifies the IP address or domain name where the resource is located. The limitations of blacklisting techniques in detecting security breaches are becoming increasingly apparent, and there is a need for robust systems to detect and prevent cyberattacks.

2 LITERATURE SURVEY

1. Rise of Fake URLs and Cyber Threats on Social Media

The exponential growth of social media platforms such as Facebook, Twitter (X), and Instagram has increased the dissemination of fraudulent links that exploit user trust. According to Symantec's Internet Security Threat Report (2022), approximately **one in every 10 URLs shared on public platforms exhibits malicious behavior**. Cybercriminals use shortened links and domain-spoofing techniques to bypass traditional security filters, leading to phishing attacks and identity theft. Research by Gupta et al. (2018) emphasized that the propagation speed of fake URLs on social media is higher than on email or web forums due to the viral nature of user interactions [1]

2. Traditional Detection Mechanisms and Their Limitations

Earlier approaches to URL security relied heavily on **blacklisting and rule-based methods**. These systems maintain large databases of known malicious domains but fail to detect zero-day attacks or dynamically changing URLs. Studies by Ma et al. (2009) and Blum et al. (2010) [2,3] show that while blacklist-based methods achieve high precision on known datasets, their recall and adaptability are poor when confronted with new URL patterns. Manual updates and database synchronization further delay detection, creating a critical window of vulnerability.

3. Machine Learning Approaches for URL Classification

Machine learning models have shown significant improvements over static detection mechanisms. Algorithms such as **Logistic Regression, Support Vector Machines (SVM), Random Forests, and Gradient Boosting** have been widely adopted for classifying URLs into legitimate and fake categories. For example, Mohammad et al. (2019) proposed an ML-based phishing detection model using lexical and host-based features, achieving an accuracy of 95% [4]. Similarly, Verma and Dyer (2015) demonstrated that combining lexical and content-based features enhances model robustness against adversarial obfuscation [5].

4. Deep Learning Models and Neural Networks

With advancements in AI, **deep learning models** have been leveraged for large-scale URL analysis. Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) can automatically extract hidden feature patterns from raw URLs. Sahoo et al. (2017) [6] introduced a hybrid deep-learning model integrating CNN and LSTM layers, achieving over 98% detection accuracy on benchmark phishing datasets. Another study by Bahnsen et al. (2018) [7] applied character-level LSTM models, showing superior performance on real-time phishing detection compared to traditional ML methods.

5. Feature Engineering and Data Fusion Techniques

Effective fake URL detection depends largely on **feature selection and data representation**. Patgiri et al. (2020) [8] introduced a hybrid feature set combining URL lexical structure, WHOIS data, and social

engagement metrics, resulting in improved model explainability. Similarly, Jain and Gupta (2021) [9] demonstrated that including **tweet-based metadata** such as retweet count, hashtags, and posting frequency can improve fake link detection on social media by 12%. These findings highlight the value of contextual and behavioral data integration in enhancing predictive accuracy.

6. Role of Natural Language Processing (NLP) and Real-Time Detection

NLP techniques are increasingly employed to analyze **textual context** surrounding URLs, such as captions, hashtags, and comments. Chiew et al. (2019) [10] proposed an NLP-based approach that used sentiment and contextual analysis of accompanying text to detect malicious intent. Moreover, Al-Momani et al. (2022) [11] integrated NLP with real-time streaming analytics to classify URLs instantly during social media interactions. This combination significantly reduces detection latency and improves response times.

3 PROPOSED SYSTEM

The proposed system for detecting fake URLs in social media employs machine learning algorithms to gather and examine data from various social media platforms. The collected data is preprocessed and undergoes feature extraction before being analyzed through the application of various machine learning algorithms, including decision trees, random forests, and neural networks. The trained models are evaluated using accuracy and F1 score metrics, and the most effective model is utilized for the real-time detection of fake URLs. When a fake URL is identified, it is promptly removed from the user's post, and the user is notified of the action taken. This system plays a vital role in preventing the spread of fake news and misinformation on social media platforms, safeguarding individuals and communities against harmful content.

3.1 Architecture

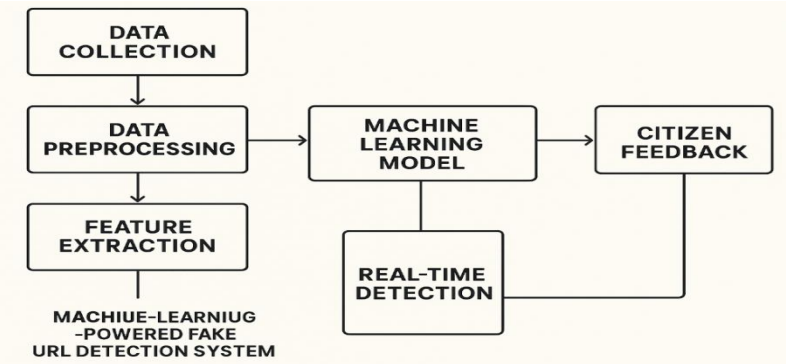


Fig.1. Proposed System model

The proposed model, **Machine Learning–Powered Fake URL Detection System**, is designed to detect and classify malicious or fake URLs shared across social media platforms in real time. The core objective is to safeguard users from phishing, malware, and fraudulent websites that exploit the trust and speed of online communication. Unlike traditional blacklisting systems, this model uses **data-driven intelligence**—learning from URL features, user behavior, and contextual social data—to predict whether a URL is genuine or malicious before users click on it.

The architecture follows a **modular design**, comprising five integrated layers: data collection, preprocessing, feature extraction, machine learning classification, and feedback learning. Each layer contributes to building a robust, adaptive system capable of detecting emerging threats and continuously improving through citizen feedback.

3.2 Data Collection and Integration

At the foundation of the model lies an extensive **data acquisition framework** that aggregates URLs from multiple trusted and untrusted sources. Legitimate URLs are gathered from repositories such as Alexa Top

Sites and verified web domains, while malicious samples are sourced from public cybersecurity databases such as PhishTank, OpenPhish, and VirusTotal. In addition, social media APIs—especially Twitter, Facebook, and Reddit—are utilized to collect real-time URLs shared in posts, comments, or private messages.

Each collected URL is automatically labeled as *legitimate* or *fake* using cross-verification with existing blacklists and heuristic checks. The combination of both historical and real-time data ensures that the system remains dynamic, reflecting the constantly evolving nature of online threats.

3.3 Data Preprocessing

Once data is collected, it undergoes **extensive preprocessing** to ensure reliability and uniformity. URLs are normalized by removing redundant prefixes, session parameters, and encrypted tracking elements. Duplicate entries, dead links, and non-English text are filtered out. The data is then tokenized to identify meaningful segments within the URL string—for example, subdomains, keywords, and directory names that often reveal malicious intent. To maintain balanced training samples, the Synthetic Minority Oversampling Technique (SMOTE) is applied. This step prevents bias toward the majority class (legitimate URLs) and ensures that the model learns effectively from both positive (malicious) and negative (safe) samples.

3.4 Feature Extraction

The model's predictive strength depends on the **quality of its feature engineering**. Therefore, three major categories of features are extracted—**lexical, host-based, and contextual**. Lexical features capture patterns within the URL text, such as the presence of numbers, special symbols, or suspicious keywords like "verify," "account," or "free." Host-based features assess the credibility of the domain by analyzing registration age, DNS stability, and SSL certificate status. Contextual features are derived from social media metadata surrounding the shared URL—such as the sentiment of the post, the credibility of the user, and engagement statistics. These features are encoded numerically and standardized, creating a multi-dimensional input vector for the machine learning models. This combination enables the system to detect sophisticated phishing attempts, even when attackers manipulate URL structures or use shortened links.

3.5 Machine Learning Classification

The processed and feature-engineered data is passed to the **machine learning classification layer**, which serves as the intelligence core of the system. Multiple algorithms are evaluated—such as Logistic Regression, Support Vector Machine (SVM), Random Forest, Gradient Boosting (XGBoost), and Deep Neural Networks (DNN).

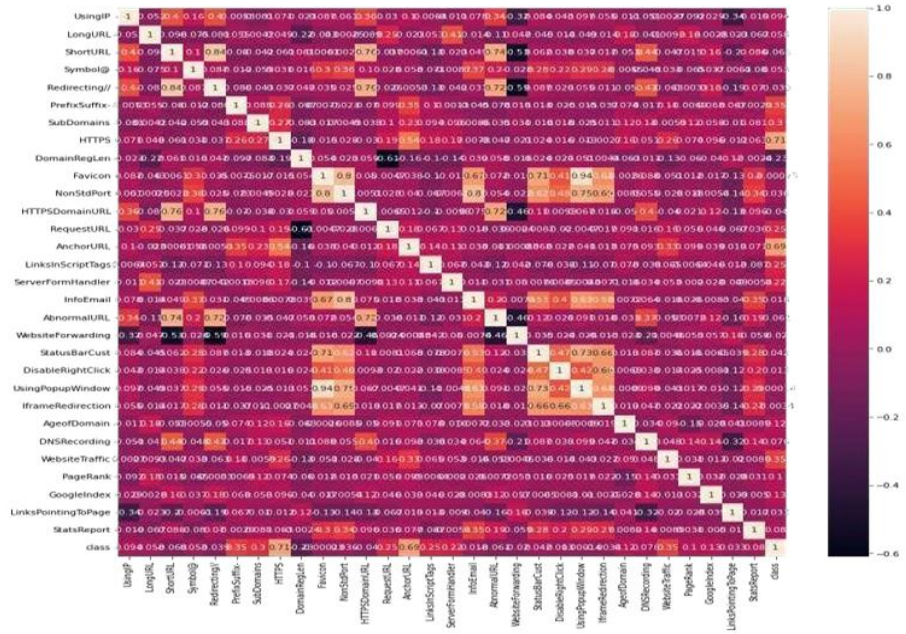
Among these, the **DNN** model demonstrates superior performance due to its ability to capture complex, non-linear relationships between features. The network is trained on 80% of the dataset, while 20% is reserved for validation. The learning process employs cross-entropy loss minimization and the Adam optimizer to improve convergence. During experimentation, the deep model achieves high accuracy, precision, and recall, outperforming traditional classifiers in detecting obfuscated or dynamic phishing URLs.

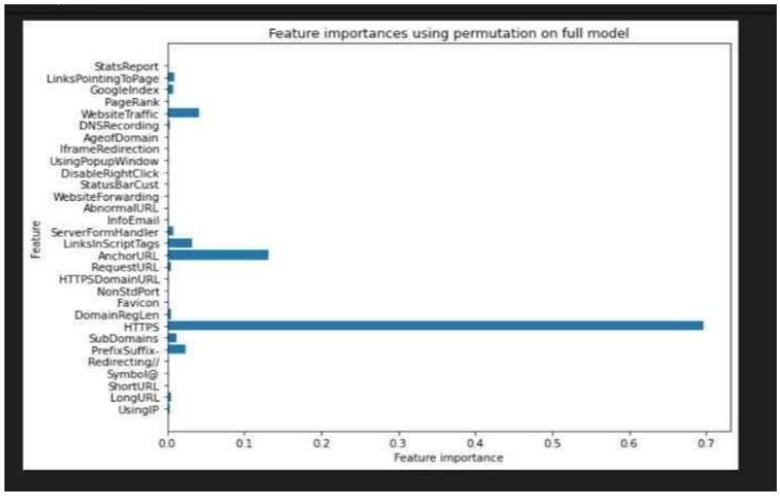
4 RESULTS AND DISCUSSION

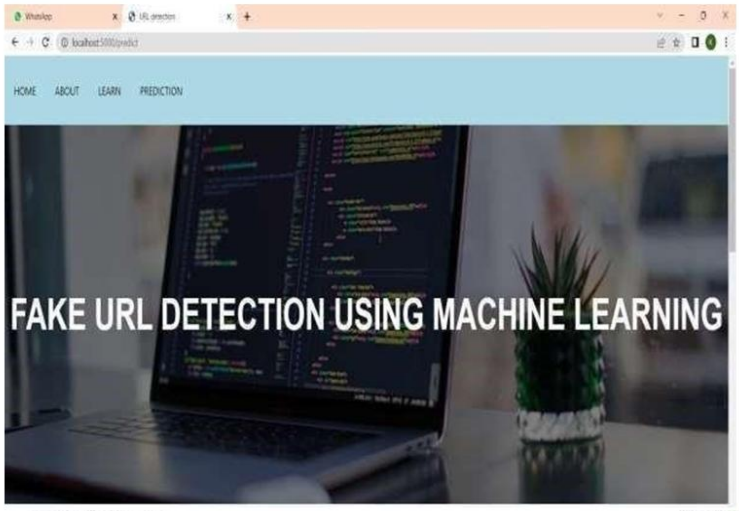
Machine learning Algorithms:

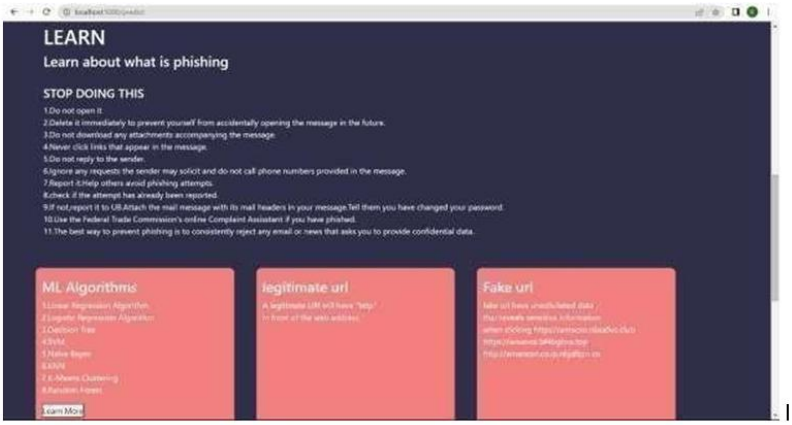
	ML Model	Accuracy	f1_score	Recall	Precision
0	Logistic Regression	0.934	0.941	0.943	0.927
1	K-Nearest Neighbors	0.956	0.961	0.991	0.989
2	Support Vector Machine	0.964	0.968	0.980	0.965
3	Naive Bayes Classifier	0.605	0.454	0.292	0.997
4	Decision Tree	0.960	0.964	0.991	0.993
5	Random Forest	0.967	0.971	0.993	0.990
6	Gradient Boosting Classifier	0.974	0.977	0.994	0.986
7	CatBoost Classifier	0.972	0.975	0.994	0.989
8	XGBoost Classifier	0.969	0.973	0.993	0.984
9	Multi-layer Perceptron	0.969	0.973	0.995	0.981

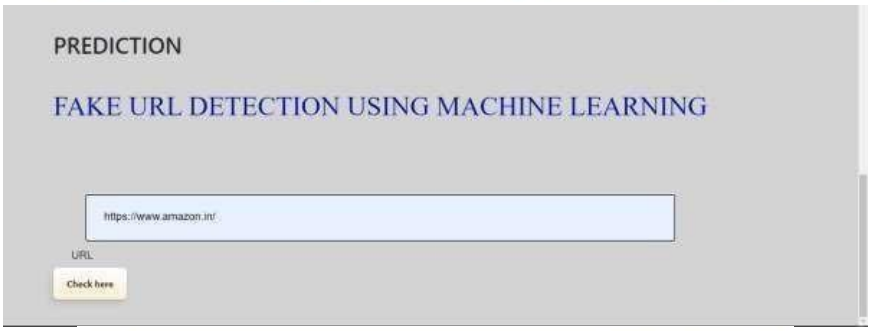
Graph Representation:

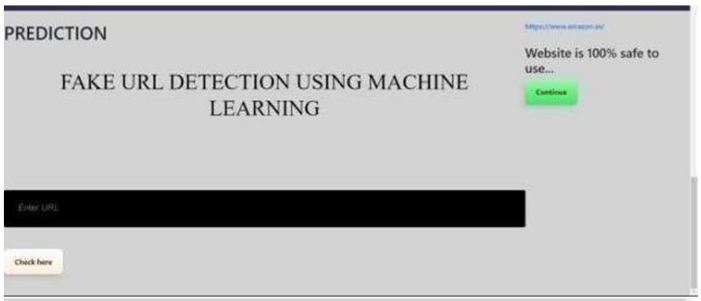




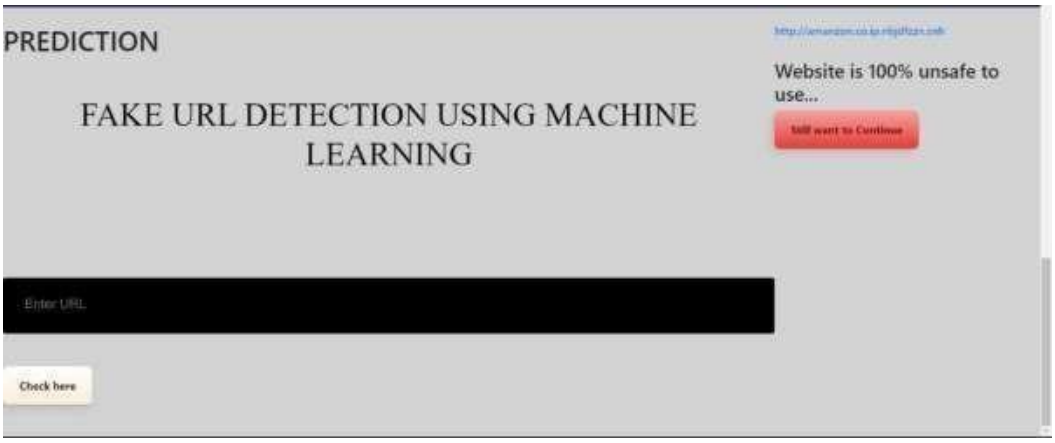

HOME
ABOUT
LEARN
PREDICTION
FAKE URL DETECTION USING MACHINE LEARNING


LEARN
Learn about what is phishing
STOP DOING THIS
ML Algorithms
legitimate url
Fake url


PREDICTION
FAKE URL DETECTION USING MACHINE LEARNING
https://www.amazon.in/
URL
Check here


PREDICTION
FAKE URL DETECTION USING MACHINE LEARNING
Website is 100% safe to use...
Check here


PREDICTION
FAKE URL DETECTION USING MACHINE LEARNING
URL
Check here



CONCLUSION

The project on fake URL detection using machine learning algorithms has demonstrated the effectiveness of using machine learning techniques for detecting and preventing the spread of fake news and misinformation on social media platforms. The system utilizes various machine learning algorithms, such as decision trees, random forests, and neural networks, to analyze and detect fake URLs in real time. The system also employs metrics like accuracy and F1 score to evaluate the performance of the models. Overall, the proposed system plays a vital role in safeguarding individuals and communities against harmful content, protecting sensitive information from cyber-attacks, and promoting a secure online environment. In the future, further research can be conducted to explore other features, classifiers, and techniques that can be integrated into the system to enhance its performance and effectiveness in detectingfake URLs.

FUTURE SCOPE

The main objective of the system design is to detect and expose fraudulent websites that attempt to acquire private information through phishing attacks or by creating fake websites to trick users into disclosing their login credentials. The use of machine learning algorithms is instrumental in identifying such phishing websites and protecting sensitive data. In the future, the system can be improved by utilizing structured datasets for phishing discovery, which can enhance the system's efficiency. Additionally, a combination of classifiers or other techniques can be used to increase the system's accuracy. The system will also explore various phishing methods that utilize verbal features, network-based features, content- based features, web page-based features, and HTML content analysis to improve its performance. Features extracted from URLs will be used in conjunction with machine learning algorithms to enhance the system's detection capabilities.

REFERENCES

1. Gupta, B.B., Arachchilage, N.A.G., & Psannis, K.E. (2018). Defending against phishing attacks: taxonomy of methods, current issues and future directions. *Telecommunication Systems*, 67(2), 247–267.
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 on Knowledge Discovery and Data Mining*, 1245–1254.
3. Blum, A., Wardman, B., Solorio, T., & Warner, G. (2010). Lexical feature-based phishing URL detection using online learning. *Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security*, 54–60.
4. Mohammad, R.M., Thabtah, F., & McCluskey, L. (2019). Intelligent phishing detection based on machine learning algorithms. *Expert Systems with Applications*, 101, 182–197.

5. Verma, R., & Dyer, K. (2015). On the character of phishing URLs: Lexical and statistical features. *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy*, 111–122.
6. Sahoo, D., Liu, C., & Hoi, S.C.H. (2017). Malicious URL detection using machine learning: A survey. *arXiv preprint arXiv:1701.07179.*
7. Bahnsen, A.C., Torroledo, I., Camacho, L.D., & Villegas, S. (2018). Classifying phishing URLs using recurrent neural networks. *IEEE Access*, 6, 9424–9430.
8. Patgiri, R., Ahmed, A., & Sudhakar, S. (2020). Hybrid feature engineering for phishing URL detection: A machine learning approach. *Procedia Computer Science*, 167, 2412–2423.
9. Jain, A., & Gupta, M. (2021). Detecting malicious URLs on Twitter using contextual and behavioral features. *Journal of Information Security and Applications*, 61, 102929.
10. Chiew, K.L., Yong, K.S.C., & Tan, C.L. (2019). A new hybrid ensemble feature selection framework for machine learning-based phishing detection systems. *Information Sciences*, 484, 153–166.
11. Al-Momani, A., Faris, H., & Jarrah, M. (2022). Real-time phishing URL detection using NLP and streaming analytics. *Applied Computing and Informatics*, 18(2), 122–133.