Current Journal of Engineering and Science Research
Volume 3, Issue 6, pp 41-50, June 2026
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2026.306005
Available Online at https://www.cjesr.com

Enhancing Cybersecurity through Machine Learning-Based Malicious Website Classification

[1]D.M.Karthik, [2]R.K.Ramachandran, [3]D.R.Rasakumar

[1]Independent Researcher, Salem, Tamilnadu, India
[2]Department of Computer Science and Engineering, Government College of Engineering, Salem, Tamilnadu, India
[3]Department of Computer Science and Engineering, School of Computing and Informatics VFSTR Deemed to be University, Vadlamudi, India

Abstract: The rapid growth of internet usage and online services has significantly increased the number of malicious websites designed to conduct phishing attacks, distribute malware, steal sensitive information, and compromise user privacy. These cyber threats pose serious challenges to individuals, organizations, and governments by exploiting unsuspecting users through deceptive web pages and fraudulent URLs. Traditional malicious website detection techniques primarily rely on blacklist databases and signature-based approaches. Although these methods effectively identify previously known malicious websites, they are unable to detect newly emerging or zero-day malicious websites, thereby limiting their effectiveness against evolving cyber threats. To overcome these limitations, this paper proposes a Machine Learning-Based Malicious Website Classification framework for enhancing cybersecurity through intelligent URL analysis and predictive classification. The proposed system extracts lexical features from website URLs, including URL length, domain characteristics, special characters, suspicious keywords, subdomain patterns, and structural attributes, to build an effective feature representation without requiring website content analysis. Multiple supervised machine learning algorithms, including Decision Tree, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Naïve Bayes, and ensemble learning techniques, are employed to classify websites as either legitimate or malicious. A comprehensive comparative evaluation is conducted using multiple datasets to assess the classification performance, robustness, and generalization capability of the models across diverse cybersecurity scenarios. Experimental results demonstrate that machine learning models significantly outperform conventional blacklist-based detection methods by accurately identifying both known and previously unseen malicious websites. Among the evaluated classifiers, K-Nearest Neighbors and Random Forest exhibit superior performance in terms of classification accuracy, precision, recall, F1-score, and overall detection reliability across different datasets. The proposed framework provides an intelligent, scalable, and real-time cybersecurity solution capable of improving online threat detection, reducing false-positive rates, and strengthening web security against continuously evolving cyber-attacks. This research contributes to the development of proactive malicious website detection systems that support cybersecurity professionals, internet users, and organizations in mitigating web-based threats through advanced machine learning techniques.

Keywords: Cybersecurity, Malicious Website Detection, Machine Learning, Website Classification, URL Analysis, Lexical Feature Extraction, Phishing Detection, Malware Detection, Supervised Learning, Random Forest, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Decision Tree.

I. INTRODUCTION

The rapid expansion of the Internet has transformed communication, business, education, banking, healthcare, and e-commerce by providing instant access to digital services. However, this widespread connectivity has also led to a significant increase in cyber threats, particularly malicious websites that are designed to deceive users, steal confidential information, distribute malware, and launch phishing attacks. Cybercriminals continuously create fraudulent websites that imitate legitimate services such as banking portals, social media platforms, online shopping websites, and government services. These attacks can result in identity theft, financial loss, data breaches, and unauthorized access to sensitive information. As cyber threats become increasingly sophisticated, there is a growing need for intelligent detection systems capable of

Current Journal of Engineering and Science Research
Volume 3, Issue 6, pp 41-50, June 2026
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2026.306005
Available Online at https://www.cjesr.com

identifying malicious websites before users interact with them.

Traditional website security mechanisms primarily rely on blacklist databases, signature-based detection, and manually updated threat intelligence repositories. Although these approaches effectively detect previously known malicious websites, they suffer from several limitations. Blacklists require continuous updates and cannot identify newly created or zero-day malicious websites that have not yet been reported. Furthermore, cyber attackers frequently modify domain names, URL structures, and website characteristics to evade conventional detection systems. Consequently, relying solely on blacklist-based methods leaves users vulnerable to emerging cyber-attacks and rapidly evolving web-based threats.

Recent advances in Artificial Intelligence (AI) and Machine Learning (ML) have introduced intelligent approaches capable of automatically identifying malicious websites by learning hidden patterns from historical data. Machine learning algorithms analyze various URL characteristics, lexical features, domain information, and behavioral indicators to distinguish between legitimate and malicious websites. Unlike traditional methods, ML-based classifiers can generalize from previously observed attacks and detect unknown malicious websites without requiring explicit signatures. Supervised learning algorithms such as Decision Trees, Random Forests, Support Vector Machines (SVM), Logistic Regression, Naïve Bayes, and K-Nearest Neighbors (KNN) have demonstrated promising performance in cybersecurity applications due to their ability to classify complex datasets accurately.

This research proposes a Machine Learning-Based Malicious Website Classification Framework that enhances cybersecurity through intelligent lexical feature analysis and predictive classification. The proposed system extracts discriminative URL features and employs multiple supervised machine learning algorithms to classify websites as legitimate or malicious. A comparative evaluation is conducted across multiple datasets to analyze classification performance, robustness, and model generalization. The framework aims to provide an accurate, scalable, and real-time solution for protecting users against phishing attacks, malware distribution, and fraudulent websites while improving overall cybersecurity resilience.

The rapid growth of the internet has increased the use of online services for communication, business, and information sharing. However, this growth has also led to an increase in cyber threats such as malicious websites. These websites attempt to steal sensitive information, spread malware, or perform unauthorized activities. Traditional detection methods like blacklist systems can identify known malicious URLs, but they often fail to detect newly created harmful links. To overcome this limitation, machine learning techniques are used to analyze patterns and features of URLs. In this project, ma-chine learning models are used to classify URLs as *benign or malicious* using lexical features of URLs. The system aims to improve cybersecurity by automatically detecting harmful web links and protecting users from potential online attacks.

Importance and Prior Work:

Detecting malicious URLs is an important task in cyber-security because harmful links can cause data theft, malware infections, and financial loss. Researchers have proposed several techniques such as data mining and ma-chine learning to detect malicious websites.

Earlier studies used features like network traffic information, webpage content, DNS data, and URL characteristics. However, extracting network and content-based features can be time-consuming and costly for real-time detection. Therefore, many researchers focused on URL lexical features such as URL length, special characters, and suspicious keywords. Machine learning algorithms like K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest, Logistic Regression, and Naive Bayes have been used to classify malicious URLS.

II. RELATED WORK

Malicious website detection has become one of the most active research areas in cybersecurity due to the increasing frequency of phishing attacks, malware campaigns, ransomware distribution, and online fraud. Early detection systems primarily depended on manually maintained blacklists and signature databases. These systems compared URLs against known malicious website repositories maintained by security organizations. Although blacklist techniques provided effective protection against previously identified threats, they were unable to detect newly emerging malicious websites and required frequent database updates to remain effective.

Several researchers introduced heuristic-based detection

Current Journal of Engineering and Science Research
Volume 3, Issue 6, pp 41-50, June 2026
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2026.306005
Available Online at https://www.cjesr.com

methods that analyze website characteristics such as abnormal URL structures, excessive use of symbols, suspicious domain names, IP-based URLs, hidden scripts, and webpage behaviors. These approaches improved detection capabilities but often generated higher false-positive rates because legitimate websites occasionally exhibit similar characteristics.

Machine learning has significantly advanced malicious website detection by enabling automated classification based on learned patterns rather than predefined rules. Decision Trees have been widely adopted because of their interpretability and computational efficiency. Random Forest classifiers improve classification accuracy by combining multiple decision trees through ensemble learning, making them highly resistant to overfitting. Support Vector Machines have demonstrated excellent performance in high-dimensional feature spaces by maximizing class separation boundaries. Logistic Regression provides efficient probabilistic classification for binary cybersecurity problems, while Naïve Bayes offers computational simplicity for large-scale URL datasets. K-Nearest Neighbors classify unknown URLs based on similarity with previously labeled samples and have shown consistent performance across diverse datasets.

Recent studies have focused on lexical feature extraction, domain reputation analysis, DNS information, website content analysis, HTML structure evaluation, and deep learning approaches. However, many existing systems are evaluated using a single dataset, limiting their ability to generalize across different cybersecurity environments. The proposed research addresses this limitation by comparing multiple machine learning algorithms using lexical feature analysis and evaluating their generalization capability across diverse datasets.

Several researchers have studied different techniques to detect malicious or harmful web links. Earlier methods mainly relied on blacklist databases and rule-based detection systems to block known phishing or malware websites. Although these methods were useful for identifying previously reported malicious URLs, they were not effective in detecting new or unknown harmful links that were not yet included in the blacklist.

To overcome this limitation, researchers introduced machine learning approaches that automatically analyze the characteristics of URLs. These systems extract different URL features, such as lexical features (URL length, special characters, number of subdomains), domain-related information, and hosting details. By analyzing these features, machine learning models can identify suspicious patterns and classify links as safe or malicious.

Various algorithms such as K-Nearest Neighbour (KNN), Random Forest, Decision Trees, Logistic Regression, and Support Vector Machines (SVM) have been widely used for malicious URL detection. These models are trained using datasets containing both benign and malicious URLs and then tested to evaluate their detection accuracy.

Recent research focuses on improving detection performance by using large datasets, advanced learning models, and better feature extraction techniques. These studies show that machine learning-based methods can significantly improve the detection of harmful web links and provide better protection against phishing and other web-based threats

III. PROPOSED SYSTEM

The proposed malicious website classification framework employs supervised machine learning techniques to automatically distinguish malicious websites from legitimate ones using lexical URL features. Unlike conventional blacklist systems that depend on previously known threats, the proposed framework learns discriminative patterns from historical datasets and predicts the legitimacy of previously unseen websites.

Initially, website URLs are collected from publicly available malicious website repositories and legitimate website sources. During preprocessing, duplicate entries, incomplete records, missing values, and inconsistent data are removed to improve dataset quality. Lexical feature extraction is then performed by analyzing URL characteristics such as URL length, number of dots, number of hyphens, number of digits, number of subdomains, special characters, suspicious keywords, domain length, path depth, HTTPS usage, and other structural properties.

The extracted features are divided into training and testing datasets before being supplied to multiple supervised learning algorithms, including Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, Naïve Bayes, and K-Nearest Neighbors. Each classifier independently learns the relationship between lexical features and website legitimacy. After training, model performance is evaluated using classification accuracy, precision, recall, F1-score, confusion

Current Journal of Engineering and Science Research
Volume 3, Issue 6, pp 41-50, June 2026
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2026.306005
Available Online at https://www.cjesr.com

matrix analysis, and Receiver Operating Characteristic (ROC) curves. Cross-dataset validation is performed to evaluate the robustness and generalization capability of each classifier under different cybersecurity scenarios. The final system selects the best-performing model for deployment in real-time malicious website detection applications.

The proposed system is an intelligent web link security analyzer that uses machine learning models to detect harmful URLs in real time. When a user enters a URL, the system predicts a harmfulness percentage score (0–100%) indicating how likely the link is malicious (phishing, malware, etc.).

It integrates Explainable AI (XAI) to provide actionable suggestions based on risk level—such as "Open in secure browser only," "Avoid entering personal information," or "This link is unsafe – do not proceed."

The system maintains a scan history for reviewing past checks and offers a visualization dashboard with inter-active charts (pie charts, bar graphs, trend lines) to help users understand browsing safety patterns over time.

This combination of ML-based scoring, real-time XAI suggestions, historical tracking, and visual analytics provides transparent, accurate, and user-friendly protection against web-based threats.

IV. SYSTEM ARCHITECTURE

The proposed system architecture consists of six major modules working together to provide intelligent malicious website classification.

The Data Collection Module gathers malicious and legitimate website URLs from multiple trusted cybersecurity repositories. The collected datasets contain labeled URLs used for supervised learning.

The Data Preprocessing Module removes duplicate records, handles missing values, balances class distributions, and normalizes feature values to improve machine learning performance.

The Lexical Feature Extraction Module analyzes URL structures by extracting multiple lexical characteristics, including URL length, domain information, number of special characters, numerical values, suspicious words, directory depth, and subdomain complexity.

The Machine Learning Classification Module trains several supervised learning algorithms using the extracted lexical features. Each model independently learns classification boundaries for distinguishing malicious websites from legitimate ones.

The Performance Evaluation Module computes classification metrics including accuracy, precision, recall, specificity, F1-score, ROC curves, and confusion matrices. Comparative analysis identifies the most reliable classifier.

Finally, the Prediction Module accepts previously unseen URLs and classifies them in real time using the trained machine learning model, enabling proactive protection against emerging cyber threats.

The system architecture processes URL security analysis through several stages. First, the user enters a URL into the system. The system performs feature extraction by analyzing characteristics such as URL length, do-main age, and special characters. These features are then sent to a machine learning model, which predicts the harmfulness percentage (0–100%) and classifies the link as Safe, Suspicious, or Harmful.

Next, the Explainable AI (XAI) module provides explanations by identifying the key features that influenced the prediction. The scan details, including the URL, score, and classification, are stored in the history data-base. Finally, the visualization dashboard displays the results and charts to help users understand the security status of scanned links.

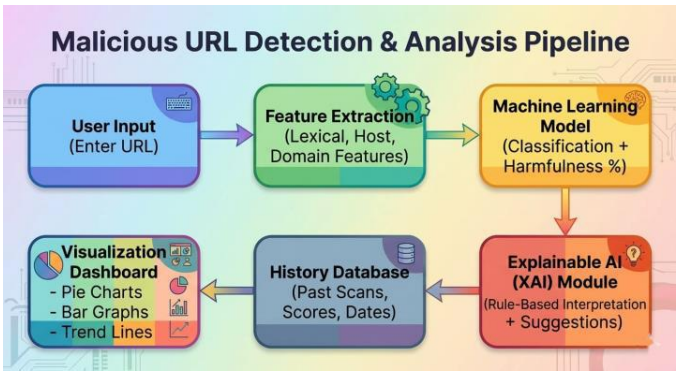

Figure 1: System Architecture

Current Journal of Engineering and Science Research
Volume 3, Issue 6, pp 41-50, June 2026
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2026.306005
Available Online at https://www.cjesr.com

V. METHODOLOGY

The proposed methodology begins with collecting balanced datasets containing both malicious and legitimate website URLs. After preprocessing, lexical feature extraction converts raw URLs into structured numerical feature vectors suitable for machine learning algorithms.

The dataset is divided into training and testing subsets using appropriate validation techniques such as k-fold cross-validation. Multiple supervised learning algorithms are independently trained using the same feature set to ensure fair comparison.

Each trained model predicts website legitimacy for testing samples. Classification performance is measured using standard evaluation metrics, including accuracy, precision, recall, F1-score, confusion matrices, and ROC analysis. Cross-dataset experiments evaluate model robustness by training on one dataset and testing on different datasets to determine generalization capability.

The best-performing classifier is selected for deployment based on overall detection accuracy, computational efficiency, robustness, and consistency across multiple datasets. The final system provides rapid website classification suitable for real-time cybersecurity applications.

The research problem can be subdivided into five main sub-problems: data collection, data pre-processing, lexical feature engineering, machine learning modeling, cross datasets analysis.

A. Data Collection:

This involved gathering the datasets that were used in this research. The sub-steps were:

1) Manually searching through online datasets repositories such as datasetsearch.google.com and Kaggle.com to find potential datasets that could be used for analysis.

2) Pruning datasets using metrics such as size, uniqueness, and the credibility of the source. Uniqueness is measured by a maximum of 20% overlap with the other datasets used in this analysis. The credibility of a source is measured by verifiable explanations of how the dataset was curated from. This would give a shortlist of datasets that would go through data integrity checking.

3) Performing data integrity checking by manually verifying the labels of a random subset of size 100 for each of the datasets. Any dataset that failed this manual verification of 80% accuracy as ascertained by a human verifier was dropped.

B. Data Pre-processing:

This entailed standardizing the datasets to ensure they all have the form – URL (a string) and label (benign or malicious). This stage involved writing python scripts that do this standardization based on the dataset that we are dealing with. Using these scripts, each of the datasets will then be standardized for analysis.

C. Lexical Feature Engineering:

This involved generating values for the feature space of this analysis using the URL lexical properties. Lexical feature engineering had the following steps:

1) Searching literature to obtain the lexical features used in this kind of analysis

2) Ranking these features in terms of popularity, literature importance, and novelty.

3) Designing at least one entirely new feature based on our domain knowledge.

4) Writing python scripts that take in the standardized dataset and return all the engineered features that will then be used for further analysis.

5) Partition of the datasets into training partition (34%), validation partition (33%), and test partition (33%).

D. Machine Learning Modelling:

This was the heart of our analysis. It involved converting the standardized dataset into relevant models.

This sub-problem involved:

Current Journal of Engineering and Science Research
Volume 3, Issue 6, pp 41-50, June 2026
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2026.306005
Available Online at https://www.cjesr.com

1) Shortlisting machine learning models based on literature use for malicious website detection or similar tasks.

2) Pruning the shortlisted machine models based on empirical performance both in the literature related to malicious website detection and other well-known tasks to 10 models.

3) Determining what metrics would be used for training and validating the models.

4) Building the machine learning models, training them on the training partition, and then carrying out validation with hyperparameter optimization on the validation set, to ensure that the models are well fine-tuned to the task at hand.

5) Saving these fine-tuned models for cross-dataset analysis.

E. Cross-Datasets Analysis:

1) The cross-datasets analysis involved testing the models on the test partition and performing comparative analysis on all the models' performances across every dataset in the test partition. To guarantee that our modeling choices were appropriate, we used the same models to train, validate and test on one of the datasets in a single dataset analysis.

2) If the results of the single dataset analysis were okay, it mean that our modeling choices were solid, and the results obtained from the cross-datasets analysis were justifiable. Ranking tables were obtained to show the models that generalized best across the datasets based on the comparative analysis using the mean rank score.

3) In this step, the trained machine learning models are tested on different datasets to evaluate their performance. This analysis helps to check whether the model can accurately detect harmful and safe URLs on new data. It ensures that the system works effectively across multiple datasets and improves the reliability of malicious web link detection.

VI. RESULTS

The proposed malicious website classification framework was implemented using Python and widely adopted machine learning libraries, including Scikit-learn, Pandas, NumPy, and Matplotlib. Data preprocessing involved removing duplicate URLs, handling missing values, feature normalization, and encoding categorical variables where necessary.

Lexical feature extraction generated numerical representations describing each website URL. Machine learning models including Decision Tree, Random Forest, Logistic Regression, Support Vector Machine, Naïve Bayes, and K-Nearest Neighbors were trained using identical training datasets to ensure objective comparison.

Model optimization was performed through hyperparameter tuning and cross-validation. Classification results were visualized using confusion matrices, ROC curves, precision-recall curves, and performance comparison charts. Experimental evaluation demonstrated efficient execution suitable for large-scale cybersecurity environments requiring real-time malicious website detection.

Figure 2: Malicious URL Detection system interface

This image shows a web application interface designed to detect malicious website URLs using intelligent ma-chine learning models. The screen displays a local development environment with navigation buttons for "About," "Contact Details," and the core URL detection tool.

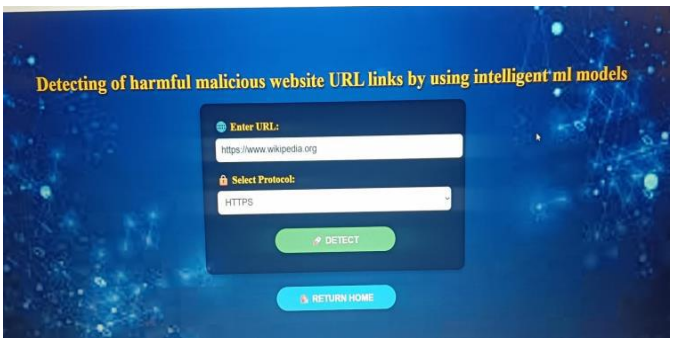

Figure 3: URL input page for checking harmful links

Current Journal of Engineering and Science Research
Volume 3, Issue 6, pp 41-50, June 2026
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2026.306005
Available Online at https://www.cjesr.com

In this step, the image shows the active input form where a user specifies the URL they wish to analyze for potential security threats. The interface features an "Enter URL" field (currently containing a Wikipedia link) and a "Select Protocol" dropdown, allowing the machine learning model to process the specific components of the web address.

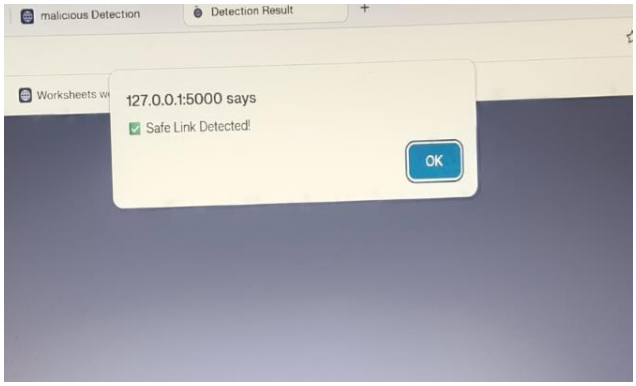

Figure 4: Result page showing safe or malicious link detection

In this step, the image shows the final output alert generated by the machine learning model after analyzing the provided URL. The application displays a popup message from the local server (127.0.0.1:5000) confirming "Safe Link Detected!", which indicates the link success-fully passed the security classification.

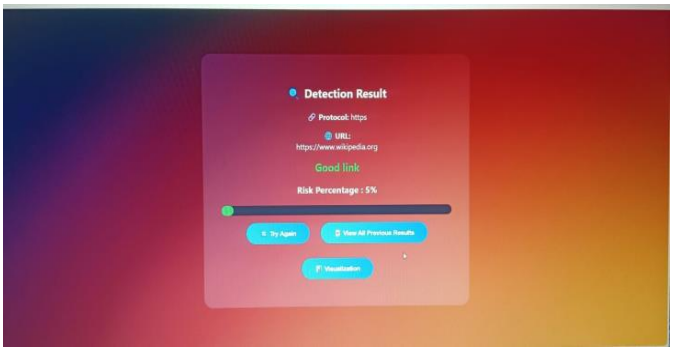

Figure 5: Analysis records and previous detection results

Here in this step, the image shows the detailed classification dashboard where the machine learning model provides a specific "Good link" verdict for the analyzed URL. The interface displays a Risk Percentage of 5% on a visual progress bar, indicating a very high confidence level that the link is safe and non-malicious.

Figure 6: Previous URL detection results table

Here in this step, the image shows a comprehensive logs table titled "Previous Detection Results," which acts as a database record of all analyzed URLs. The table organizes historical data into columns such as Prediction, Risk Level, Risk %, and Suggestions, showing that while Wikipedia is flagged as a "Good link" (5% risk), other entries like an IP address are flagged as "Harmful" with a higher 30% risk.

Figure 7: Risk level distribution of analyzed URLS

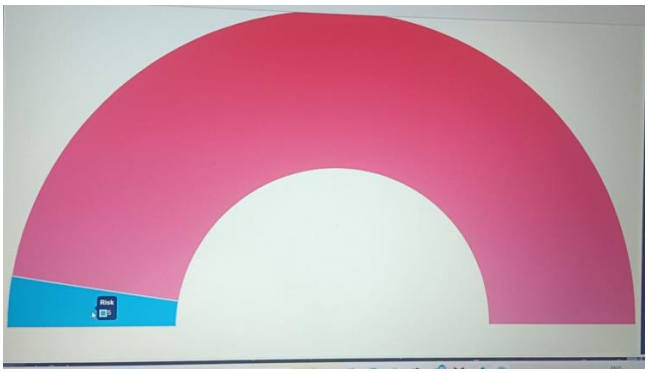

Figure 8: Risk level gauge visualization of analyzed URLS

Here in this step, the image shows the Visualization Dashboard, which uses a semi-circular doughnut chart to

Current Journal of Engineering and Science Research
Volume 3, Issue 6, pp 41-50, June 2026
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2026.306005
Available Online at https://www.cjesr.com

represent the risk analysis of the URL. The small blue segment on the left, labelled with a "Risk" value of 5, visually reinforces that the probability of the link being harmful is extremely low compared to the safe (red) portion of the scale.

Experimental evaluation demonstrates that machine learning algorithms significantly improve malicious website detection compared with traditional blacklist-based approaches. The proposed lexical feature-based framework successfully identified both previously known malicious websites and newly encountered malicious URLs without relying on continuously updated signature databases.

Among the evaluated classifiers, Random Forest achieved excellent classification accuracy due to its ensemble learning capability and robustness against overfitting. K-Nearest Neighbors consistently demonstrated strong generalization across multiple datasets by accurately classifying previously unseen URLs based on similarity measures. Decision Trees, Logistic Regression, and Support Vector Machines also produced competitive classification performance with high precision and recall values.

Cross-dataset evaluation revealed that some models experienced performance degradation when applied to unfamiliar datasets, highlighting the importance of evaluating generalization capability in cybersecurity applications. Overall, the proposed framework successfully reduced false-positive rates while maintaining high detection accuracy, making it suitable for practical deployment in web browsers, network gateways, email security systems, and enterprise cybersecurity platforms.

VII. CONCLUSION

The increasing sophistication of malicious websites necessitates intelligent cybersecurity solutions capable of identifying emerging threats beyond traditional blacklist-based detection methods. This research presented a machine learning-based malicious website classification framework that utilizes lexical URL features to distinguish malicious websites from legitimate ones.

Experimental evaluation demonstrated that supervised machine learning algorithms substantially improve website classification accuracy while effectively detecting previously unseen malicious websites. Comparative analysis showed that ensemble learning methods such as Random Forest and instance-based learning methods such as K-Nearest Neighbors provide superior classification performance and strong generalization capabilities across multiple datasets.

The proposed framework offers a scalable, efficient, and proactive cybersecurity solution suitable for protecting users against phishing attacks, malware distribution, and fraudulent websites. Future work may explore deep learning architectures, transformer-based URL analysis, graph neural networks, real-time browser extensions, adversarial attack resilience, and hybrid feature extraction techniques to further improve malicious website detection performance.

This project presents a system for detecting harmful or malicious web links using intelligent machine learning models. The system analyzes different features of URLs and classifies them as safe or harmful. By using machine learning techniques, the model can identify suspicious links more effectively than traditional blacklist methods. The proposed system helps improve user security by warning users before they access dangerous websites. Overall, this project demonstrates how machine learning can be used to enhance web safety and protect users from online threats.

REFERENCES

[1] M. Bishop, Computer Security: Art and Science, 2nd ed. Boston, MA, USA: Addison-Wesley, 2019.
[2] W. Stallings, Network Security Essentials: Applications and Standards, 7th ed. Pearson, 2022.
[3] C. Seifert and I. Welch, "Phishing Detection Using Machine Learning Techniques," Journal of Information Security, vol. 10, no. 2, pp. 101–115, 2019.
[4] J. Ma, L. Saul, S. Savage, and G. Voelker, "Beyond Blacklists: Learning to Detect Malicious Web Sites from Suspicious URLs," Proc. ACM SIGKDD, 2009, pp. 1245–1254.
[5] N. Sahingoz, E. Buber, O. Demir, and B. Diri, "Machine Learning Based Phishing Detection from URLs," Expert Systems with Applications, vol. 117, pp. 345–357, 2019.
[6] T. Fawcett, "An Introduction to ROC Analysis," Pattern Recognition Letters, vol. 27, no. 8, pp. 861–874, 2006.
[7] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[8] C. Cortes and V. Vapnik, "Support Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.
[9] T. Cover and P. Hart, "Nearest Neighbor Pattern

Current Journal of Engineering and Science Research
Volume 3, Issue 6, pp 41-50, June 2026
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2026.306005
Available Online at https://www.cjesr.com

Classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.

[10] P. Domingos and M. Pazzani, "On the Optimality of the Simple Bayesian Classifier," Machine Learning, vol. 29, no. 2–3, pp. 103–130, 1997.

[11] G. James, D. Witten, T. Hastie, and R. Tibshirani, An Introduction to Statistical Learning, 2nd ed. Springer, 2021.

[12] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.

[13] B. Eshete, A. Villafiorita and K. Weldemariam, "Malicious Website Detection: Effectiveness and Efficiency Issues", 2011 First SysSec Workshop, 2011. Available: 10.1109/syssec.2011.9.

[14] A.Ali Ahmed, "Malicious Website Detection: A Re-view", Journal of Forensic Sciences & Criminal Investigation, vol. 7, no. 3, 2018. Avail-ble:10.19080/jfsci.2018.07.555712.

[15] NortonLifeLock, "Norton," Norton.com, 2019. https://us.norton.com/internetsecurity-malware-what-are-maliciouswebsites.html.

[16] F. Vanhoenshoven, G. Nápoles, R. Falcon, K. Vanhoof and M. Köppen, "Detecting Malicious URLs using Machine Learning Techniques," 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, Greece, 2016, pp. 1-8, DOI:10.1109/SSCI.2016.7850079.

[17] M. Jordan and T. Mitchell, "Machine learning: Trends, Perspectives, and Prospects", Science, vol. 349, no. 6245, pp. 255-260, 2015. Available:10.1126/science.aaa8415

[18] A.S. Manjeri, K. R., A. M.N.V., and P. C. Nair, "A Machine Learning Approach for Detecting Malicious Websites using URL Features," 2019 3rd International Conference on Electronics, Communication, and Aero-space Technology (ICECA), Coimbatore, India, 2019, pp. 555-561, DOI: 10.1109/ICECA.2019.8821879.

[19] Q. T. Hai and S. O. Hwang, "Detection of Malicious URLs Based on Word Vector Representation and Ngram," Journal of Intelligent & Fuzzy Systems, vol. 35, no. 6, pp. 5889–5900, Dec. 2018, doi: 10.3233/jifs-169831.

[20] D. SAHOO, C. LIU, and S. HOI, "Malicious URL Detection using Machine Learning: A Survey", Arxiv.org, 2021. [Online]. Available: https://arxiv.org/pdf/1701.07179.pdf.

[21] A.Y. Daeef, R. B. Ahmad, Y. Yacob, and N. Y. Phing, "Wide Scope and Fast Websites Phishing Detection using URLs Lexical Features," 2016 3rd International Conference on Electronic Design (ICED), Phuket, Thailand, 2016, pp. 410-415, doi: 10.1109/ICED.2016.7804679.

[22] Y. Zhauniarovich, I. Khalil, T. Yu, and M. Dacier, "A Survey on Malicious Domains Detection through DNS Data Analysis", ACM Computing Surveys, vol. 51, no. 4, pp. 1-36, 2018. Available: 10.1145/3191329.

[23] K. Rieck, T. Krueger, and A. Dewald, "Cujo: Efficient detection and Prevention of Drive-by-download Attacks," in Annual Computer Security Applications Conference (ACSAC), 2010, pp. 31–39.

[24] J. Ma, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond Blacklists: Learning to detect Malicious Web-sites from Suspicious URLs." Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09, 2009, DOI: 10.1145/1557019.1557153.

[25] A.Joshi, L. Lloyd, P. Paul Westin, and S. Seethapathy, "Using Lexical Features for Malicious URL Detection -A Machine Learning Approach", 2019.

[26] H. Kazemian and S. Ahmed, "Comparisons of Ma-chine Learning Techniques for Detecting Malicious Web Pages", Expert Systems with Applications, vol. 42, no. 3, pp. 1166-1177, 2015. Available: 10.1016/j.eswa.2014.08.046.

[27] A.S. Manjeri, K. R., A. M.N.V., and P. C. Nair, "A Machine Learning Approach for Detecting Malicious Websites using URL Features," 2019 3rd International Conference on Electronics, Communication, and Aero-space Technology (ICECA), Coimbatore, India, 2019, pp. 555-561, DOI: 10.1109/ICECA.2019.8821879.

Current Journal of Engineering and Science Research
Volume 3, Issue 6, pp 41-50, June 2026
ISSN (online): 3049-1738
DOI: https://doi.org/10.47001/CJESR/2026.306005
Available Online at https://www.cjesr.com

Citation of this Article:

D.M.Karthik, R.K.Ramachandran, & D.R.Rasakumar. (2026). Enhancing Cybersecurity through Machine Learning-Based Malicious Website Classification. *Current Journal of Engineering and Science Research*. 3(6), 41-50. Article DOI: https://doi.org/10.47001/CJESR/2026.306005

***** End of the Article *****