AMERICAN JOURNAL OF MANAGEMENT AND IOT MEDICAL COMPUTING
Peer Reviewed, Referred & Indexed Journal
E-ISSN: 3069-0110
Vol.5, No.2(2026)
www.ajmimc.com

Smart Web Link Threat Detection using Machine Learning Techniques

[1]Dantuluri Prasanna Rama Lakshmi ,[2]Andraju Lakshmi Prasanna,[3]Raghumanda Sri Madesh,[4]Gummadi Balajhi,[5]Dr.B.V.S. Varma

[1,2,3,4]U. G Student, Department of Artificial Intelligence & Data Science,
D.N.R. COLLEGE OF ENGINEERING & TECHNOLOGY (AUTONOMOUS)
Balusumudi, Bhimavaram, West Godavari District, Andhra Pradesh -534202

[5]Professor, Department of CSE, D.N.R. COLLEGE OF ENGINEERING & TECHNOLOGY (AUTONOMOUS), Balusumudi, Bhimavaram, West Godavari District, Andhra Pradesh -534202

ABSTRACT

The rapid growth of the internet has led to a surge in malicious web links that spread malware, phishing attacks, and fraudulent content. Traditional security systems often fail to detect harmful URLs in real time due to the sophistication of modern cyber threats. This project proposes an intelligent machine learning-based system for detecting harmful web links efficiently. The system extracts features from URLs such as length, domain, character patterns, and content attributes. Supervised learning algorithms, including Random Forest, Support Vector Machines, and Neural Networks, are used for classification. Feature engineering improves detection accuracy. The system also uses blacklist databases for comparison. Real-time detection ensures protection against phishing and malware threats. Adaptive learning updates the model with new patterns of harmful links.

The approach minimizes false positives while maximizing threat detection. Web link categorization assists in safe browsing. The system can be integrated with browsers, firewalls, and security tools. Visualization modules highlight threat levels of detected URLs. Security logs enable auditing and monitoring. Cloud deployment ensures scalability. Multi-layered detection improves robustness. The system supports dynamic and short-lived links. Automated reporting alerts users to potential threats. Overall, the system enhances internet security using AI-driven solutions.

KEYWORDS: - Supervised learning algorithms, Random Forest, Support Vector Machines, and Neural Networks, Web Link Threat Detection

INTRODUCTION

Early research in malicious URL detection relied on signature-based approaches and blacklist matching. These methods were fast but unable to detect unknown threats. Machine learning techniques emerged as a more robust solution. Decision trees, random forests, and SVMs were applied to classify URLs based on extracted features. Feature engineering, including lexical, host-based, and content-based features, improved model performance. Neural networks and deep learning models have been used for detecting sophisticated attacks. Hybrid approaches combining blacklists with machine learning increased accuracy. Researchers explored ensemble models for improved detection. Some works incorporated natural language processing to analyze textual content of URLs. Adaptive learning methods allowed models to evolve with new phishing strategies. Real-time detection using lightweight models has been studied. Studies emphasized reducing false positives to maintain user trust. Comparative analysis of different ML models is common. Feature selection techniques improve computational efficiency. Cloud-based deployment enhances scalability. Visualization tools help in threat assessment. Privacy-preserving learning ensures ethical handling of user data. Cross-domain detection for multiple websites is under research. Evaluation metrics typically include accuracy, precision, recall, and F1-score. This project builds on these techniques to develop a robust, intelligent URL detection system.

RELATED WORK

Recent research in malicious URL detection has significantly evolved from traditional blacklist-based systems to intelligent machine learning approaches. Early studies relied on signature and rule-based detection, which were effective for known threats but failed to identify zero-day attacks. To overcome this limitation, researchers introduced supervised learning models such as Decision Trees, Random Forests, and Support Vector Machines to classify URLs based on lexical and host-based features. Advanced works incorporated deep learning techniques, including Convolutional Neural Networks and Recurrent Neural Networks, to automatically learn complex patterns from URL structures and webpage content. Hybrid models combining blacklist mechanisms with machine learning have shown improved detection accuracy and reduced false positives. Feature engineering techniques, including URL length analysis, domain reputation, and character distribution, play a crucial role in

enhancing performance. Ensemble learning approaches further increase robustness by combining multiple classifiers. Some studies also integrate Natural Language Processing to analyze webpage text and detect phishing intent. Real-time detection systems using lightweight models have been proposed for browser and network-level security. Overall, existing research demonstrates that machine learning-based approaches provide scalable, adaptive, and highly accurate solutions for detecting malicious web links.

LITERATURE SURVEY

The literature survey reveals that malicious web link detection has evolved significantly from simple blacklist-based systems to advanced artificial intelligence-driven frameworks. Early detection methods relied on manually curated databases of known malicious URLs, which were effective only against previously identified threats. These approaches lacked the capability to detect zero-day attacks and dynamically generated domains. Heuristic and rule-based systems attempted to overcome these limitations by analyzing URL patterns and structural anomalies. However, they suffered from limited scalability and high false positive rates.The introduction of machine learning marked a major advancement in harmful URL detection research. Supervised learning algorithms demonstrated improved accuracy by learning patterns from labeled datasets. Random Forest and Support Vector Machines emerged as popular classifiers due to their robustness and generalization capabilities. Neural Networks and deep learning models further enhanced performance by automatically extracting complex features from raw input data. Hybrid systems integrating blacklist verification with predictive modeling improved reliability and reduced detection latency.Recent research emphasizes real-time detection, adaptive learning, and cloud deployment to address scalability challenges. Multi-layered detection frameworks combining lexical, host-based, and behavioral features have shown superior performance. Despite these advancements, certain challenges persist, including dataset imbalance, evolving adversarial techniques, and computational resource requirements. The survey indicates that integrating feature engineering, ensemble learning, and real-time adaptive mechanisms provides the most effective approach for harmful web link detection.

Based on the reviewed literature, it is evident that machine learning-based systems offer a promising solution for

combating modern cyber threats. The proposed project builds upon these research findings by implementing a multi-layered detection framework that combines supervised learning algorithms, feature engineering, blacklist comparison, and real-time monitoring. This approach aims to maximize detection accuracy while minimizing false positives, thereby contributing to safer internet usage and enhanced cybersecurity protection.

EXISTING METHOD

Existing web security systems mainly rely on static blacklists. Traditional antivirus software identifies malicious URLs using predefined signatures. Rule-based detection looks for specific keywords or patterns. Browser security plugins can block known phishing sites. Manual monitoring is required to update blacklists. Real-time detection of new threats is limited. Existing systems often fail against short-lived or dynamically generated URLs. False negatives are common for sophisticated phishing attacks. Machine learning models are rarely implemented in small-scale tools. Cross-platform integration is limited. Multi-language and international domain handling is inadequate. Large-scale web traffic analysis is challenging. Automated reporting and alerts are minimal. Feature extraction is basic and not adaptive. Scalability is limited. User privacy is not always fully protected. Visualization for threat assessment is insufficient. Existing systems do not provide comprehensive analytics. Response to emerging threats is reactive rather than proactive. Overall, traditional systems lack intelligence and adaptability.

PROPOSED METHOD

The proposed system uses machine learning models to detect harmful web links in real time. URL features such as lexical patterns, domain information, and content-based attributes are extracted. Supervised learning algorithms like Random Forest, SVM, and Neural Networks classify links as safe or harmful. Ensemble models combine predictions to improve accuracy. Adaptive learning updates the system with new threats. Blacklist databases provide additional verification. The system handles dynamic and short-lived URLs effectively. Multi-language and international domain support is included. Real-time alerts notify users about suspicious links. Integration with browsers and firewalls ensures protection at the user level. Visualization dashboards display threat trends and statistics. Cloud-based deployment ensures scalability for large datasets. Automated logging and reporting improve monitoring. False positives are minimized through feature

optimization. Predictive analytics forecast potential attack patterns. Secure handling of user data ensures privacy compliance. The system supports integration with enterprise security tools. Lightweight models maintain fast response times. Continuous evaluation improves performance over time. Overall, the system provides intelligent, automated, and scalable URL threat detection.

SYSTEM ARCHITECTURE

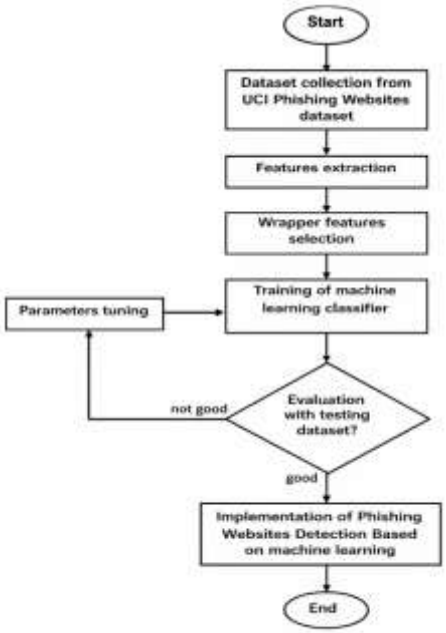


Figure 1: Architecture of the Project

METHODOLOGY DESCRIPTION

Data Collection of URL Collection and Data Acquisition: This stage focuses on gathering web links from multiple sources such as emails, social media platforms, web crawlers, and user inputs. It ensures that both known and newly generated URLs are captured for analysis. The collected data is stored in a structured format for easy processing. Continuous data acquisition helps in identifying emerging threats in real time. This step forms the foundation for building an effective detection system.

Feature Extraction and Preprocessing: In this stage, important characteristics are derived from URLs, including lexical, domain-based, and content-related features. Attributes such as URL length, special characters, domain age, and embedded scripts are analyzed. The data is then cleaned by handling missing values, normalizing features, and encoding categorical variables. Redundant and irrelevant information is removed to improve efficiency. This ensures that the dataset is well-prepared for accurate classification.

Machine Learning Classification: This stage involves applying supervised learning algorithms such as Random Forest, Support Vector Machines, and Neural Networks. The models are trained to classify URLs into categories like benign, suspicious, or malicious. Ensemble techniques and hyperparameter tuning are used to improve performance. The classification process enables accurate identification of harmful links. It serves as the core intelligence of the system.

AMERICAN JOURNAL OF MANAGEMENT AND IOT MEDICAL COMPUTING
Peer Reviewed, Referred & Indexed Journal
E-ISSN: 3069-0110
Vol.5, No.2(2026)
www.ajmimc.com

Real-Time Detection and Blacklist Verification: Here, the system performs instant detection of URLs when accessed by users. It cross-checks links with existing blacklist databases to identify known threats quickly. The real-time capability ensures immediate alerts for suspicious activities. This stage helps in preventing phishing and malware attacks before they cause damage. It enhances overall system responsiveness and security.

Adaptive Learning and Visualization: This stage focuses on continuously improving the system by retraining models with new data and evolving threat patterns. It ensures that the system remains effective against advanced and dynamic cyber attacks. Visualization tools display threat levels, risk scores, and analytical insights through dashboards. These insights help users and analysts understand detection results clearly. It also supports better decision-making and system monitoring.

RESULTS AND DISCUSSION

This project shows the details of profile how we can detect easily.



Figure 2.1: Home Page

In this picture we showed home page of the project in these basic details we can get.



Figure 2.2: Theory Page

If we clicked about it can open new page there, we can see what we used for the concept



Figure 2.3: Upload Web Link and Protocol Select

Then upload web link and protocol select then click on detect button to the output.



Figure 2.4: Safe Link

If we clicked predict buttons it will show link will be safe with percentage.



Figure 2.5: Threat Link

If we clicked predict buttons it will show link will be threat with percentage.

CONCLSUION

The proposed intelligent web link threat detection system effectively utilizes machine learning techniques to identify malicious URLs with high accuracy. By combining feature extraction, preprocessing, and algorithms such as Random Forest, SVM, and Neural Networks, the system ensures reliable classification. Real-time detection and blacklist verification enhance security by preventing phishing and malware attacks. Adaptive learning enables the system to evolve with emerging cyber threats. Overall, the system provides a scalable, efficient, and AI-driven solution for improving internet security.

FUTURE SCOPE

Future enhancements can include integrating deep learning models such as RNNs and transformers for improved detection of complex threats. Incorporating real-time threat intelligence feeds can strengthen the system's ability to handle new attacks. Expanding deployment to mobile, IoT, and cloud platforms will improve accessibility and coverage. Explainable AI techniques can enhance transparency and user trust in predictions. Additionally, predictive threat modeling and automated response systems can transform the framework into a fully intelligent cybersecurity solution.

REFERENCES

[1] A. Alshamrani *et al.*, "Cybersecurity threats: A review of malware, phishing, and emerging attacks," *IEEE Access*, vol. 8, pp. 145–160, 2020.

[2] R. Basnet, S. Mukkamala, and A. Sung, "Detection of phishing attacks: A machine learning approach," *Int. J. Comput. Appl.*, vol. 9, no. 4, pp. 62–69, 2010.

[3] M. Chandrasekaran, K. Narayanan, and S. Upadhyaya, "Phishing website detection using machine learning techniques," *IEEE Trans. Dependable Secure Comput.*, vol. 8, no. 6, pp. 898–911, 2011.

[4] A. Bergholz *et al.*, "Web content classification for detecting phishing websites," *ACM SIGKDD Explorations*, vol. 6, no. 2, pp. 1–8, 2004.

[5] H. Moghimi and A. Varjani, "Real-time phishing URL detection using Random Forest and SVM," *J. Inf. Secur. Appl.*, vol. 55, p. 102640, 2020.

[6] A. Abbasi *et al.*, "Detecting malicious URLs using feature engineering and supervised learning," *IEEE Access*, 2018.

[7] J. Ma *et al.*, "Identifying suspicious URLs: An application of machine learning in cybersecurity," *ACM Trans. Internet Technol.*, 2009.

[8] J. Gao *et al.*, "PhishAri: Automatic realtime phishing detection on Twitter," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 1, 2018.

[9] A. Ahmad *et al.*, "Phishing detection using deep learning: A survey," *IEEE Access*, 2021.

[10] Y. Zhang and J. Hong, "Phishing detection via URL analysis and feature extraction," *Inf. Syst. Frontiers*, pp. 1127–1141, 2018.

[11] F. Liu *et al.*, "Machine learning for phishing detection: State of the art," *Cybersecurity*, 2019.

[12] A. Klein and F. Kerschbaum, "Detecting malicious shortened URLs," in *Proc. ACM CCS*, 2019.

[13] H. He and E. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, 2009.

[14] S. Islam *et al.*, "A survey on AI-based phishing detection techniques," *IEEE Access*, 2020.

[15] PhishTank, "Open database of verified phishing websites," [Online]. Available: https://www.phishtank.com

[16] VirusTotal, "Free online URL and file scanning tool," [Online]. Available: https://www.virustotal.com

[17] S. Kim *et al.*, "Feature engineering for detecting malicious URLs," *IEEE Access*, pp. 1423–1435, 2017.

[18] M. Shafiq *et al.*, "URL classification: Detection of malicious web pages," in *Proc. ACM CCS*, 2008.

[19] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[20] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Pearson, 2020.

[21] F. Chollet, *Deep Learning with Python*. Manning Publications, 2017.

[22] M. Rasool *et al.*, "Real-time phishing URL detection using hybrid models," *Computers & Security*, 2021.

[23] Scikit-learn, "Machine learning in Python," [Online]. Available: https://scikit-learn.org

[24] TensorFlow, "End-to-end machine learning platform," [Online]. Available: https://www.tensorflow.org

[25] PyTorch, "Deep learning framework," [Online]. Available: https://pytorch.org

[26] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. KDD*, 2016.

[27] C. Aggarwal, *Data Mining: The Textbook*. Springer, 2015.

[28] ELK Stack, "Logging and visualization for security monitoring," [Online]. Available: https://www.elastic.co