International Journal of Engineering Research and Science & Technology

Identification Of Malicious URL's Using Machine Learning for Proactive Cyber Threat Prevention

[1]Dr. A. Tirupatiah,[2]Kasukurthi Divyajyothi,[3]Gandikota Pavani,[4]Borugadda Arunkumar

[1]Associate Professor, Dept of Computer Science and Engineering, St. Ann's College of Engineering and Technology, Chirala-523187, India.

[2,3,4]B. Tech Student, Dept of Computer Science and Engineering, St. Ann's College of Engineering and Technology, Chirala-523187, India.

ABSTRACT

The rapid growth of internet usage has significantly increased cyber threats, especially through malicious URLs. These URLs are commonly used in phishing, malware distribution, and fraudulent activities. Traditional blacklist-based detection methods fail to identify newly generated malicious links. This project proposes a machine learning-based approach to identify malicious URLs proactively. The system extracts meaningful features from URLs and applies machine learning models to classify them as safe or malicious. Advanced algorithms such as XGBoost improve detection accuracy by learning complex URL patterns. Additionally, the system provides an evidence summary to explain why a URL is classified as malicious. This approach enhances user awareness and strengthens cybersecurity defences.

Keywords: ***Malicious URLs, Machine Learning, Cyber Security, XG-Boost, Phishing Detection ,Feature Extraction, URL Classification***

INTRODUCTION

The internet has become an essential platform for communication, business, and information sharing. However, this growth has also led to an increase in cyber attacks, particularly those involving malicious URLs, Attackers use deceptive links to steal sensitive information, spread malware, or redirect users to fake websites. Existing security mechanisms are often ineffective against newly created malicious URLs. Machine learning provides a dynamic solution by learning patterns from historical data and identifying unknown threats. This project focuses on detecting malicious URLs using supervised machine learning techniques. By converting URLs into numerical features and training efficient models, the system aims to provide accurate and proactive threat detection. The

proposed solution enhances online safety by preventing threat detection. The proposed solution enhances online safety by preventing users from accessing harmful websites.

LITERATURE REVIEW

Several studies have explored malicious URL detection using machine learning. Researchers have applied algorithms such as Naive Bayes, Support Vector Machines, and Random Forest for URL classification. Some approaches rely on lexical features like URL length and special characters, While others use host-based and content-based features. Although these methods achieve reasonable accuracy, they often suffer from high false- positive rates and limited adaptability. Recent studies suggest that esemble and boosting techniques provide better performance. However, many existing systems do not offer explainability for predictions. This project address these gaps by using XGBoost and providing an evidence-based explanation for classification results.

RELATED WORK

Related work in malicious URL detection mainly focuses on traditional machine learning and blacklist approaches. Rule-based systems depend heavily on predefined patterns and fail against zero-day attacks. Some research integrates deep learning models, but these require large datasets and high computational resources. Other studies emphasize feature engineering but lack real-time usability. Compared to existing works, the proposed system balances accuracy, efficiently, and explainability. It combines feature-based learning with rule-based evidence generation to improve user trust and system transparency.

EXISTING METHOD

Existing systems for malicious URL detection primarily rey on blacklists and signature-based techniques. These methods can only detect known malicious URLs and fail to identify newly generated threats. Manual rule-based approaches exist, but they often use limited features and simple classifiers resulting in lower accuracy. Additionally, most existing systems do not provide reasons for their predictions, reducing user confidence. These limitations highlight the need for a more robust and explainable solution.

PROPOSED METHOD

The proposed system uses machine learning techniques to identify malicious URLs proactively. URLs provided by users are preprocessed and converted into numerical features such as length, special characters, protocol type, and keyword presence. These features are fed into advanced machine

learning models, particularly XGBoost, to perform classification. The system not only determines whether a URL is safe or malicious but also generates an evidence summary explaining the decision. This improves accuracy, adaptability, and transparency. The proposed approach effectively mitigates cyber threats in real time.

ARCHITECTURE

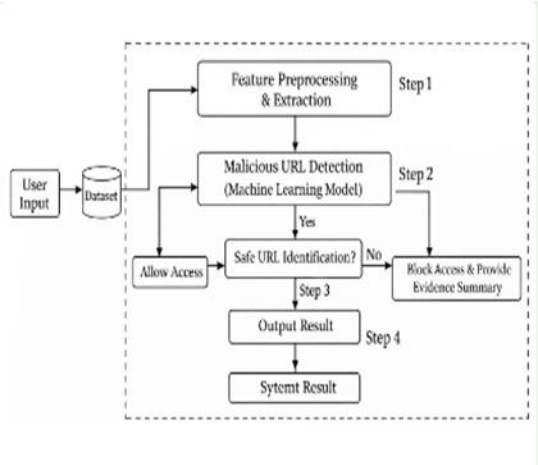

Fig 1: System design

METHODOLOGY DESCRIPTION

The methodology begins with dataset collection containing both safe and malicious URLs. Data preprocessing is performed to clean and normalize the URLs. Feature extraction converts URLs into meaningful numerical attributes . The dataset is split into training and testing sets. Multiple machine learning models such as Logistic Regression, Random Forest, and XGBoost are trained and evaluated. XGBoost is selected due to its superior performance. The trained model is integrated into the system for real-time prediction. When a user inputs a URL, the system process it through the same feature pipeline and produces the final classification along with evidence. This step-by-step approach ensures accuracy and reliability

RESULT AND DISCUSSION

HOME PAGE:

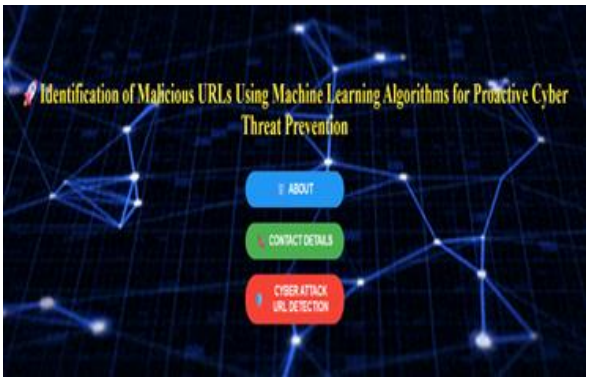

Fig 2: Home Page

The home page provides easy navigation with options such as About, Contact, and URL Detection for user interaction.

Fig 3: About Page

The About Page explains the purpose of the system and highlights its role in detecting malicious URLs using machine learning.

Fig 4: Contact Page

The Contact page allows users to reach the administrator for queries, feedback or support related to the system.

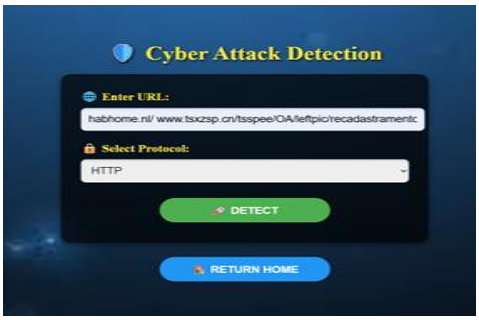

Fig 5: URL Detection Page

This Page allows users to enter a URL, which is analyzed by the system to determine whether it is safe or malicious.

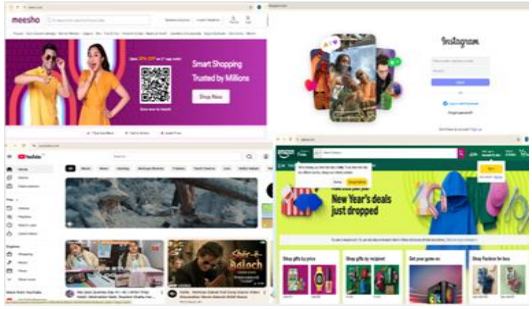

Fig 6: Safe Results

If the URL is safe, the system allows the user to directly access the requested webpage.

Fig 6: Malicious Result

CITIZEN REGISTRATION PAGE:

If the URL is malicious, the system blocks access and displays an evidence summary explaining the reason for detection.

CONCLUSION

This Project presents an efficient machine learning-based system for detecting malicious URLs. By extracting important URL features and applying classification models, the system accurately identifies unsafe links. The proposed approach helps in preventing cyber-attacks such as phishing and malware distribution. Overall, the system improves user security and trust while browsing the internet.

FUTURE SCOPE

In the future, the proposed system can be enhanced by integrating deep learning models to further improve detection

accuracy. The system can be developed as a real-time browser extension or mobile application for instant URL verification. Safe URLs can be highlighted in blue color, while malicious URLs can be displayed in red color along with alert messages to warn users. Continues model updates using live internet data can help in detecting newly emerging threats. Additionally, the system can be extended to identify malicious link in emails, text messages, and QR codes. These enhancements will improve usability, security, and real-time threat prevention.

REFERENCES

1. Ramachandran, V., Kumari, Y. S., & Harini, P. (2016). Image retrieval system with user relevance feedback. *Computer Science Engineering, St. Anns College of Engineering, Chirala*.
2. J. Ma et al., "Detecting malicious websites using machine learning," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2009, pp. 1045–1054.
3. S. Garera et al., "A framework for detection of phishing attacks using URL features," in Proc. ACM Workshop on Recurring Malcode (WORM), 2007, pp. 1–10.
4. D. Sahoo, C. Liu, and S. C. Hoi, "Malicious URL detection using machine learning: A survey," ACM Computing Surveys, vol. 50, no. 4, pp. 1–36, 2017.
5. H. Le et al., "URLNet: Learning a URL representation with deep learning for malicious URL detection," arXiv preprint arXiv:1802.03162, 2018.
6. X. Chen et al., "Malicious URL detection using supervised learning techniques," International Journal of Security and Its Applications, vol. 10, no. 1, pp. 1–14, 2016.
7. R. Verma and A. Das, "What's in a URL: Fast feature extraction and malicious URL detection," in Proc. IEEE Conf. Communications and Network Security (CNS), 2017, pp. 1–9.
8. M. Al-Saleh et al., "Machine learning techniques for phishing detection: Review and evaluation," Computers & Security, vol. 76, pp. 123–136, 2018.
9. M. Zouina and B. Outtaj, "A lightweight phishing detection scheme using SVM," International Journal of Computer Applications, vol. 170, no. 2, pp. 15–20, 2017.
10. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794.
11. Kaggle, "Malicious and benign URL dataset," Kaggle Repository. [Online]. Available: https://www.kaggle.com