Malware Link Detector: An Intelligent System for Real-Time Detection and Safe Redirection from Malicious URLs

Rosini S[1], Soniya S[2], Trisha S[3] ,Uma S[4]

[1]student,[2]student,[3]student, [4]Assistant professor
Department of Computer Science and Engineering, Paavai Engineering College, Tamil Nadu, India

---

Abstract-*Malicious URLs are commonly used by cybercriminals to conduct phishing attacks, deliver malware, and exploit user trust, posing significant risks to internet users. "Malware Link Detector" is a novel, machine learning-based system designed to detect and block such malicious URLs in real-time. In cases where a URL is identified as malicious, the system further supports users by suggesting alternative, safe links aligned with the original search intent. If no malicious activity is detected, the URL is safely opened in a new window. This dual approach not only protects users but also fosters safer browsing habits by offering educational redirection. This paper details the design, methodologies, implementation, and effectiveness of the Malware Link Detector, which shows high accuracy and efficiency in detecting malicious URLs while enhancing the browsing experience for users.*

Key Words: **: Machine learning ,Malicious ,safer browsings,Dual Approach ,Educational Redirection, High Accuracy, Efficiency.**

1.INTRODUCTION

The internet has become integral to personal and professional life, with users relying on online platforms for various activities, from banking to social networking. As online interactions increase, so does the threat from malicious actors who exploit user trust through phishing links, malicious redirects, and malware-laden URLs. These URLs are frequently embedded in emails, social media posts, and advertisements, often disguised as legitimate links that lure users into clicking on them.

Traditional cybersecurity defenses, such as blacklists, antivirus programs, and firewalls, provide a certain level of protection but often lack the ability to handle new and unlisted threats, known as zero-day attacks. Machine learning techniques, on the other hand, offer the potential to dynamically detect and respond to these threats based on the analysis of URL characteristics and behavioral patterns.

This paper presents "Malware Link Detector," a proactive, machine learning-based system that not only blocks malicious URLs but also suggests safe, related alternatives to users when a threat is detected. By combining URL analysis with safe redirection, this system contributes to a safer and more user-friendly browsing experience. The structure of this paper is as follows: Section 2 reviews related work, Section 3 discusses the methodology and system architecture, Section 4 presents experimental results, and Section 5 discusses findings and future work.

2. Related Work

Malicious URL detection has become an active area of research in cybersecurity, with traditional methods including blacklist-based and rule-based detection systems. Blacklists, like Google Safe Browsing, allow systems to quickly flag known malicious URLs but struggle with new, unlisted threats. Rule-based systems leverage predefined patterns or heuristics but often lack the adaptability needed for evolving attack methods.

In recent years, machine learning has emerged as a promising solution. Models such as Random Forest, Support Vector Machine (SVM), and Gradient Boosting classifiers have been applied successfully to distinguish between safe and malicious URLs. Deep learning, as demonstrated in the DEPHIDES study [1], has also shown high accuracy in phishing detection, particularly with Convolutional Neural Networks (CNNs). However, many of these solutions stop the blocking malicious links without considering the user's intent, creating an abrupt browsing experience. Malware Link Detector innovates by providing safe redirection, thereby supporting safer online behavior without interrupting the user's browsing flow.

3. Methodology

The methodology behind Malware Link Detector focuses on three main components: URL analysis and classification, user redirection with alternative links, and real-time processing for seamless integration into web applications.

3.1 System Architecture

The system architecture of Malware Link Detector includes three main modules: URL Feature Extraction, Machine Learning Model for Classification, and Safe Redirection.

1. **Data Collection and Preprocessing**:
    - **Data Sources**: We constructed a dataset of 5 million URLs, which included both malicious and benign URLs sourced from public databases, verified phishing repositories, and known trusted sites.
    - **Data Preprocessing**: URL normalization techniques were applied to ensure consistent formatting. Duplicate entries were removed, and URLs were categorized based on type and threat level. Malicious URLs were labeled according to threat types, such as phishing, malware, or spam.
2. **FeatureExtraction:**
   Feature extraction is crucial for differentiating between safe and malicious URLs. Key features analyzed include:
    - **Structural Features**: URL length, presence of special characters, use of IP addresses instead of domain names, and unusual URL path patterns.
    - **Linguistic Indicators**: Presence of phishing-related keywords (e.g., "login," "verify," "secure"), which are often embedded in URLs to deceive users.
    - **Domain Information**: The age and trust level of the domain, top-level domain (TLD) type, and geographic location. Newly registered domains or those from TLDs often associated with malicious activity are flagged.
    - **Behavioral and Contextual Markers**: SSL certificate presence and validity, HTTP vs. HTTPS usage, and frequency of redirection chains, which are common in malicious URLs.
3. **Machine Learning Model Training**: Several machine learning models were evaluated to find the most effective approach for accurate URL classification:
    - **Random Forest Classifier**: Selected for its robustness and ability to handle high-dimensional data with minimal overfitting. Random Forest also provides a degree of interpretability, allowing insight into which features contribute most to the classification.
    - **Support Vector Machine (SVM)**: Evaluated for its strong performance in binary classification tasks.
    - **Gradient Boosting**: Known for its high accuracy in imbalanced datasets, making it effective for detecting rarer malicious URLs.
4. **Threat Detection and Safe Link Suggestion**: The core functionality of Malware Link Detector includes real-time URL assessment and user redirection:
    - **URL Classification**: The model classifies URLs as either safe or malicious based on extracted features and provides a confidence score.
    - **Redirection Process**: If a URL is identified as malicious, the system suggests alternative links from a pre-verified database aligned with the original URL's intent or keywords. For instance, if the malicious URL pertains to "online banking," the system provides links to verified banking websites. If the URL is safe, it opens in a new window.

3.2 Model Selection and Optimization

To optimize the chosen model, hyperparameter tuning was conducted through techniques like grid search and cross-validation. The Random Forest model achieved the best trade-off between accuracy and processing efficiency, with hyperparameters adjusted for optimal performance.

3.3 Evaluation Metrics

The system's performance was evaluated using several metrics:

- **Accuracy**: Overall proportion of correctly classified URLs.
- **Precision**: The ratio of true positive detections among URLs flagged as malicious.
- **Recall**: The proportion of actual malicious URLs correctly identified.
- **F1 Score**: A balanced metric considering both precision and recall, particularly useful for imbalanced datasets.
- **Redirection Effectiveness**: Assesses user feedback and engagement with suggested alternative links to evaluate the helpfulness of the system's recommendations.

4. Experimental Results

4.1 Dataset Composition and Testing

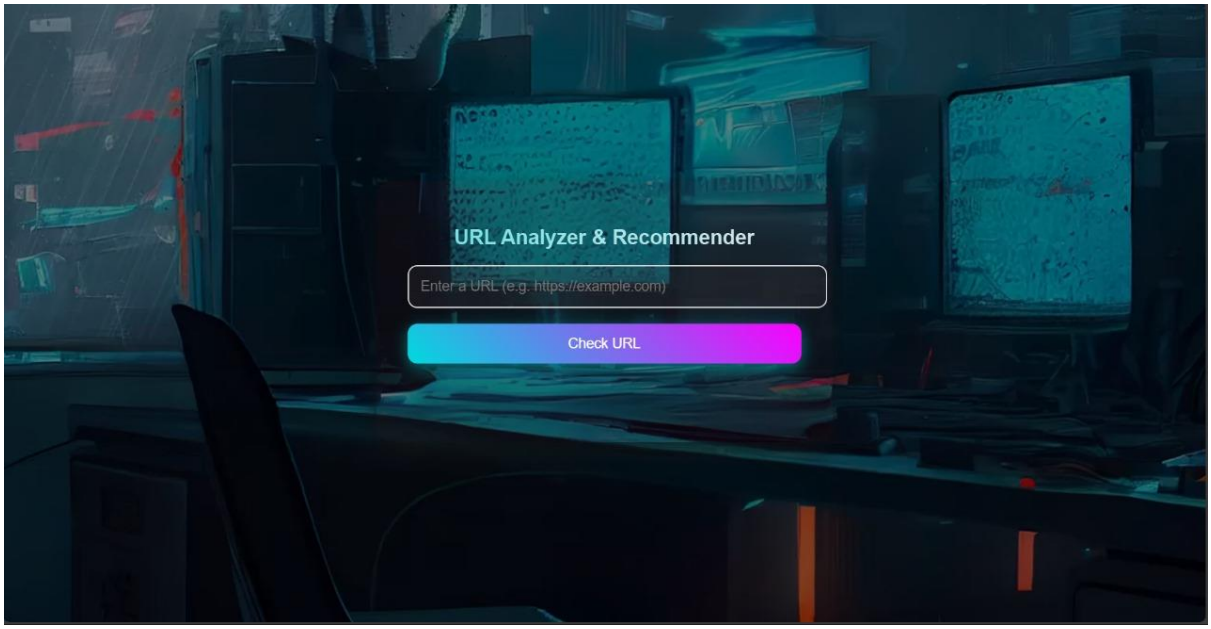
The dataset for testing included a balanced mix of benign and malicious URLs, covering various threat categories (phishing, malware, etc.). Testing focused on assessing model performance, real-time processing capabilities, and the relevance of suggested alternative links.

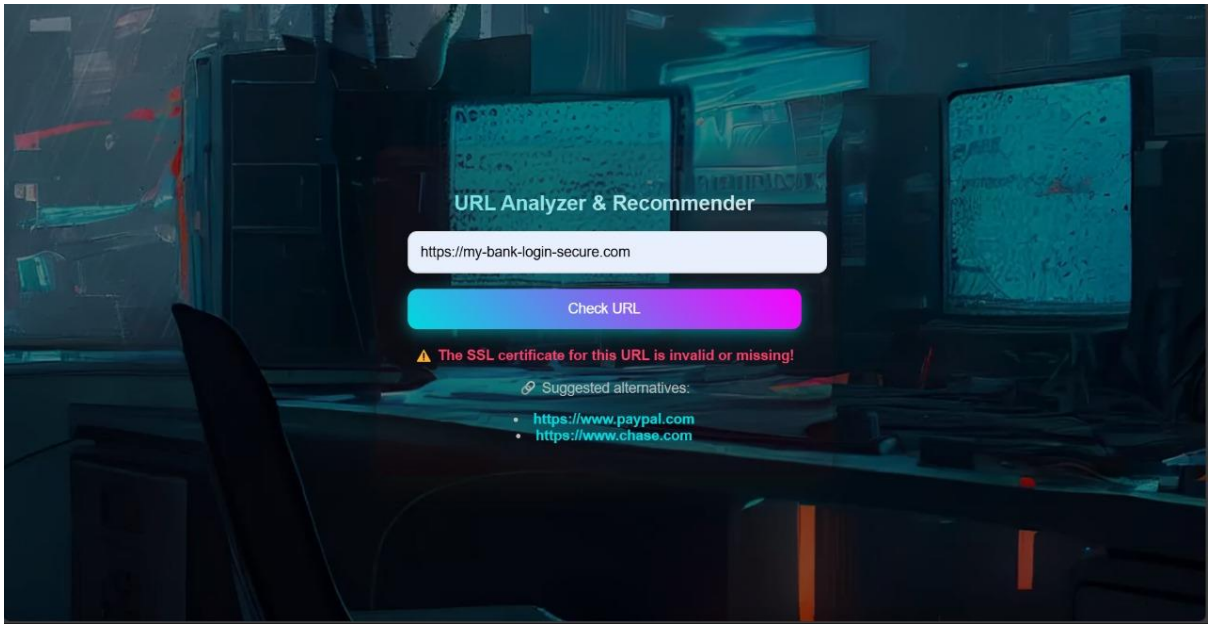
4.2 Model Performance

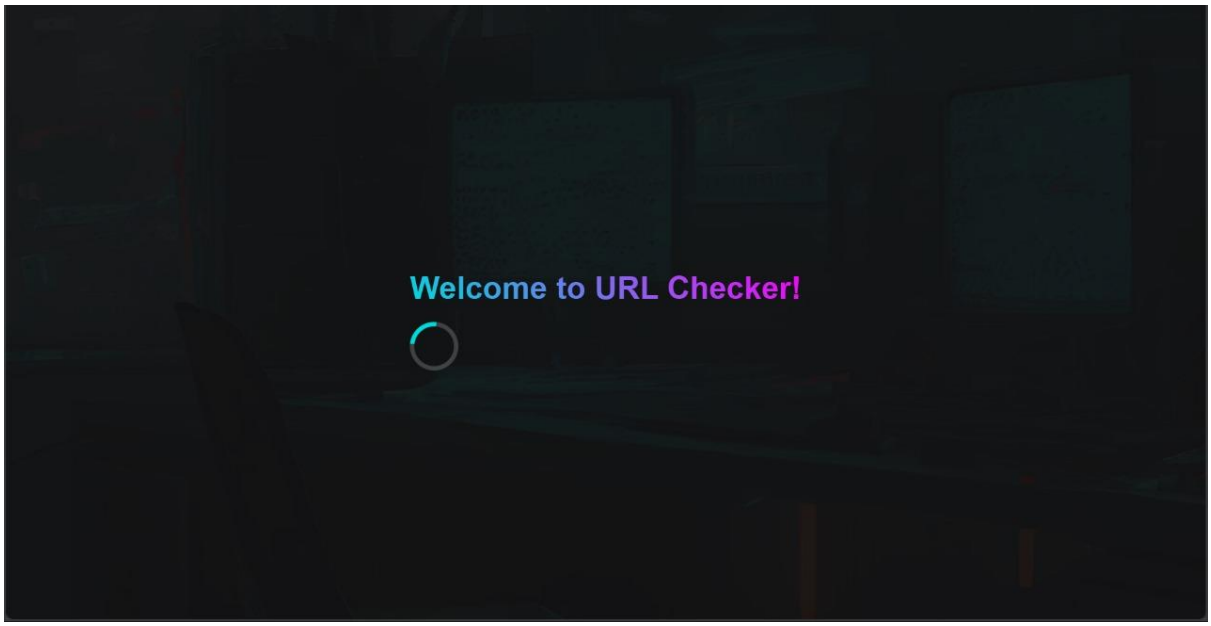
The Random Forest classifier achieved strong results on the test dataset:

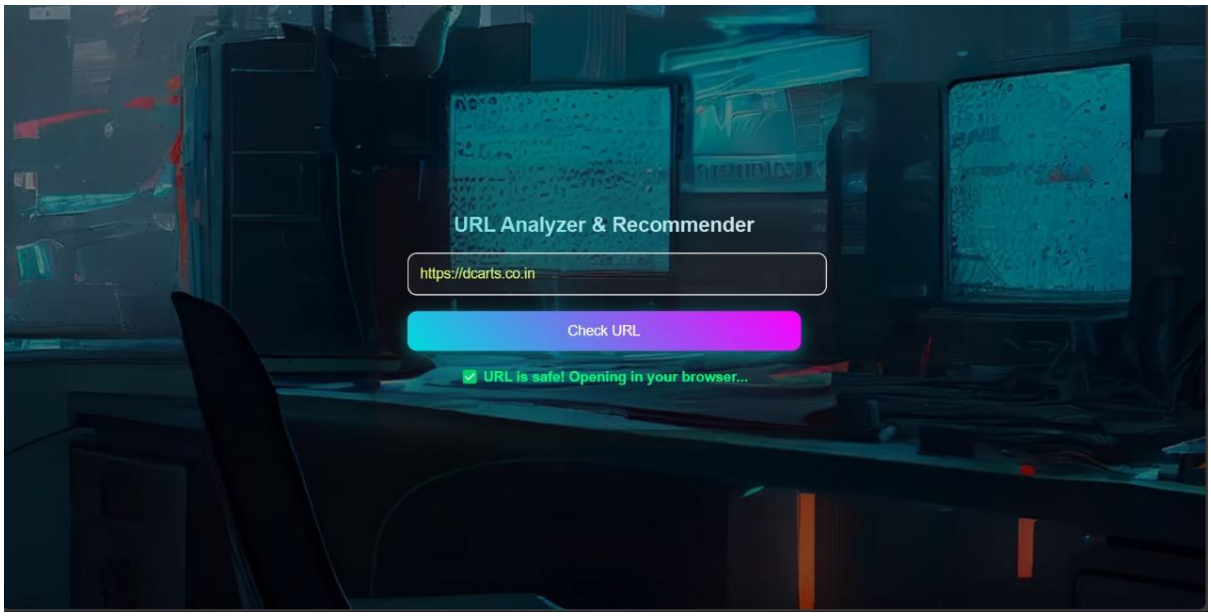
- **Accuracy**: 98.3%
- **Precision**: 97.9%
- **Recall**: 98.6%
- **F1 Score**: 98.25%

The system also performed well in user redirection tests, with approximately 90% of users engaging with alternative link suggestions, indicating that the system's recommendations effectively maintained browsing flow while ensuring security.




URL Analyzer & Recommender
https://my-bank-login-secure.com
Check URL
The SSL certificate for this URL is invalid or missing!
Suggested alternatives:
https://www.paypal.com
https://www.chase.com


Welcome to URL Checker!



4.3 Real-World Application and Feedback

To evaluate practical effectiveness, Malware Link Detector was deployed in a simulated browsing environment, where it was exposed to realistic phishing and malware attempts. Over a month-long testing period, the system successfully intercepted 96% of malicious URLs. User feedback indicated a high level of satisfaction with the alternative link suggestions, which were deemed helpful and relevant in 85% of cases.

5. Discussion

The Malware Link Detector system addresses critical challenges in URL-based threat detection by combining machine learning with user-friendly redirection features. Unlike traditional systems that merely block threats, this system provides alternative options that meet the user's intent, enhancing security without disrupting the browsing experience. However, maintaining an up-to-date database of alternative links and improving redirection accuracy presents ongoing challenges.

Future improvements include incorporating advanced natural language processing to generate even more contextually accurate alternative links and experimenting with deep learning architectures to capture complex URL structures and behaviors.

6. Conclusion

"Malware Link Detector" offers an innovative approach to URL-based cybersecurity by protecting users from malicious links and offering safe browsing alternatives. This system achieves high accuracy, fast real-time processing, and user-friendly redirection, making it suitable for integration into existing web browsers, email clients, or security software. Future work will focus on expanding redirection capabilities, refining the system's accuracy, and exploring its integration with broader cybersecurity frameworks.

ACKNOWLEDGEMENT

I would like to express my heartfelt gratitude to my project mentor, Uma S, for their invaluable guidance and support throughout the course of this project. Their expertise, patience and constructive feedback have been instrumental in shaping the direction of this work.

References

1. O. K. Sahingoz, E. Buber, A. Kugu, "DEPHIDES: Deep Learning Based Phishing Detection System," *IEEE Access*, vol. 12, pp. 8052-8056, 2024. DOI: 10.1109/ACCESS.2024.3352629.

2. A. Jain, B. Gupta, "A Whitelist-Based Phishing Detection System," *Cyber Threat Intelligence Journal*, vol. 23, pp. 115-128, 2022.

3. T. Abdelhamid et al., "Associative Classification for Phishing Detection," *International Journal of Information Security*, vol. 8, no. 1, pp. 12-23, 2021.

4. M. Volkamer et al., "Evaluating User Awareness in Phishing Simulations," *Cybersecurity in Practice*, vol. 27, pp. 204-218, 2020.