International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 05 | May - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Suspicious Link Analyzer Using Artificial Intelligence

Mrs B. Mamatha

G. Anjali
K. Kiran Kumar
T. Shravan Kumar

Abstract:

With the rapid expansion of internet usage and the increased sharing of hyperlinks through emails, messages, and social media platforms, the spread of malicious or suspicious links has become a major cybersecurity concern. These links can lead to phishing websites, malware downloads, or data theft. This paper proposes an AI-driven Suspicious Link Analyzer (SLA) system that can intelligently classify and detect suspicious or harmful URLs based on their features. The proposed model uses machine learning techniques to identify patterns in URLs and predict potential threats. This paper discusses the architecture, dataset, feature extraction, and the performance of various classification algorithms in detecting suspicious links.

1. Introduction

The internet has become an essential tool in modern communication, commerce, and education. However, with its benefits come significant risks, including cybercrime. One of the most common vectors of cyberattacks is through malicious or suspicious links, often sent via emails or posted on social media platforms. These links may appear legitimate but redirect users to phishing websites or initiate malware downloads. Manual detection is time-consuming and inefficient. Thus, automated systems powered by artificial intelligence (AI) are crucial for efficient detection and prevention.

2. Problem Statement

Suspicious and malicious links pose a serious threat to internet users. Traditional blacklisting methods are not scalable or effective in identifying new and evolving threats. Hence, there is a need for an intelligent system that can detect and analyze suspicious URLs in real time, based on their inherent characteristics rather than relying solely on pre-existing databases.

3. Literature Review

Previous works have utilized various techniques such as heuristic-based methods, signature detection, and blacklisting to detect malicious URLs. However, these methods lack adaptability. Recent studies have employed machine learning (ML) models like Decision Trees, Random Forests, Support Vector Machines (SVM), and Deep Learning algorithms to detect suspicious links. These models have shown improved accuracy but still require optimization in terms of speed, feature extraction, and real-time prediction.

4. Methodology

4.1 Data Collection

The dataset comprises URLs labeled as "benign" or "malicious/suspicious," sourced from online repositories such as PhishTank, VirusTotal, and OpenPhish.

4.2 Feature Extraction

Key features extracted from URLs include:

- URL length
- Presence of IP address in the URL
- Number of special characters (e.g., '@', '-', '_', '.')
- Use of HTTPS protocol
- Domain age and reputation
- Number of subdomains
- Use of URL shortening services

4.3 Model Architecture

The system uses supervised learning models trained on labeled URL datasets. The models considered include:

- Logistic Regression
- Random Forest
- Naive Bayes
- XGBoost
- Neural Networks (for deep learning approach)

4.4 Workflow

1. Input URL is pre-processed.
2. Features are extracted from the URL.
3. The trained ML model analyzes the features.
4. The output classifies the URL as either "benign" or "suspicious."

5. Results and Discussion

Experimental results show that Random Forest and XGBoost models performed best, with an accuracy of over 95% on the test dataset. The models effectively detected obfuscated or phishing URLs that traditional methods missed. The deep learning approach, though accurate, required higher computational resources and was slower compared to ensemble models.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Reg.	88%	85%	83%	84%
Random Forest	96%	94%	95%	94%
Naive Bayes	85%	80%	78%	79%
XGBoost	97%	96%	95%	95.5%
Neural Network	95%	94%	92%	93%

6. Conclusion

The study demonstrates that AI-based models, particularly ensemble methods like Random Forest and XGBoost, are effective in analyzing and classifying suspicious links. By leveraging features from URLs and applying machine learning, the system can offer real-time predictions, thereby enhancing user safety. This Suspicious Link Analyzer can be integrated into browsers, email clients, or antivirus software to provide proactive protection against cyber threats.

7. Future Work

Future improvements could include incorporating real-time web content analysis, natural language processing of surrounding text, and active URL behavior tracking. Integration with threat intelligence platforms can further improve accuracy.

References

1. Ma, J., Saul, L. K., Savage, S., & Voelker, G. M. (2009). Beyond blacklists: learning to detect malicious web sites from suspicious URLs.

2. Mohammad, R. M., Thabtah, F., & McCluskey, L. (2012). Intelligent phishing detection system using association rule mining.

3. Le, T., Markopoulou, A., & Faloutsos, M. (2011). Phish Def: URL names say it all.

4. URL Haus, Phish Tank, Virus Total datasets (accessed 2024).