IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

MALWARE AND MALICIOUS WEBSITE DETECTION; A DEEP LEARNING APPROACH

A Chrome Extension for Malicious URL and Malware Download Prevention using LSTM and Neural Networks

[1]Mr. Abhishek Narvekar, [2]Mr. Yash sawant, [3]Mr. Saurabh Kamble, [4]Asst. Prof. T. V. Deokar
[1,2,3]Student, [4]Asststant Professor
[1,2,3,4]Department of Data Science,
[1,2,3,4]D.Y. Patil College of Engineering and Technology, Kolhapur, India

Abstract: Malicious websites or uniform resource locator (URLs) are a huge concern in the field of cyber-security. It has always proven to be a threat to society to access such malicious websites that result in comprising the system. Many cases have occurred where malicious websites have penetrated user's computer and their privacy was compromised. These websites are a host to various cyber tools that are used to remotely control or corrupt a device. The cyber tools used are spams, malwares, trojan horses and many more. This has resulted in various losses throughout the world not only financial but also emotional. This has made these malicious urls a global threat. Traditional classification methods include blacklists, periodic reporting and signature comparisons based on data volume, changes, processes over time and relationships between features. In our project we have tried to use the deep learning approach to detect and block these websites. Also, as mentioned visiting such malicious websites result compromises the device's security, downloading malware containing files is also a huge concern. Many websites that are pirated are available on the internet which might contain downloadable malware files. Therefore, there is a need to block such downloads before they are downloaded and attack the system. The Malwares are responsible for corruption of file and this results in compromising the user's privacy as well as his financial loss. To eliminate this problem, we are trying to develop a extension that is trained as a deep learning model that blocks the download before completion and helps the user to stay away from malwares.

***Index Terms* - Malicious URL detection, Deep learning, LSTM, Malware detection, Chrome extension, Cybersecurity, Static feature analysis, Threat prevention.**

1) INTRODUCTION

Malicious User Resource Locators (URL) are websites that are harmful to access and remotely control a system with the owner's consent. They can harm a device in many ways for example, installing a spyware, or an encrypting the user's data and demanding ransom (ransomware). All these concerns have put forward the need to secure a device from such malicious websites. URLs are component of websites that can be used to identify the websites if they are safe or malicious. URL are the address of a resource available on world wide web(WWW). The URL can be broken in to 2 components. One is Rule. Various types of Rules like hypertext transfer protocol are example of rules. Various protocols that are used to identify and sort websites. These protocols are FTP, DNS, etc. The second is Resource identifier. It is the webpage's address used to locate the resource. Just like a file on your computer can be found by giving it a file name, a URL are used to find desired website. URL is the address of a WWW resource. A service provider is the address of a web page on the Internet. Unsuspecting users visit such sites and become victims of various types of scams, including loss of money, theft of personal information (identity, credit card, etc.), and installation of malware. Phishing is a major attack that is a result of malicious websites. Drive-by downloads are the (unintentional) downloading of malware when accessing a URL. These attacks are usually carried out by exploiting vulnerabilities in plugins or by injecting malicious code via JavaScript. Malware is malicious code that is injected into legal services to commit crimes. Most free downloadable software is a major source of malware. This includes free software such as games, web browsers, free anti-virus software, etc. Malwares have a defined structure and hence they can be classified into various categories. Malwares mostly are used to manipulate a person into paying to a lump amount of money to retrieve their data. Malwares are used by hacker by coating them under downloadable files which penetrate a device and then attack the system. To prevent cyber systems from operating major important divisions, researchers have studies to develop various malware stoppers.

There are a bunch of tools and anti-viruses like QuickHeal available in the market to detect malware and network attacks, but they all have limitations. Anti-viruses work against malware signatures, and the signature information should be changed regularly.

2) Literature Review

The paper "Malicious URL Detection: A Comparative Study"[1] by Maheshwari, Shantanu, Janet, B., and Kumar R. investigates the structural patterns of URLs to identify malicious websites. The authors analyze URL formats, including domain names, paths and query parameters, to distinguish between benign and harmful sites. Their approach relies on feature extraction from URL syntax rather than content analysis, making it efficient for real-time detection. The study compares heuristic-based methods with statistical models, finding that hybrid approaches yield better accuracy. However, the study acknowledges limitations in detecting dynamically generated or obfuscated URLs, which can evade traditional pattern-matching techniques. Future enhancements could include integrating machine learning to improve adaptability against evolving phishing techniques, as well as incorporating real-time threat intelligence feeds for up-to-date detection.

The survey "Malicious URL Detection using Machine Learning: A Survey"[2] by Steven C.H. Hoi, Doyen Sahoo, and Chenghao Liu explores various machine learning techniques for classifying URLs as malicious or benign. The study emphasizes binary classification models, evaluating their performance in detecting phishing, malware distribution, and scam websites. The authors compare traditional algorithms like Logistic Regression and Decision Trees with advanced methods such as XGBoost and Neural Networks, highlighting the latter's superior performance in handling large-scale datasets. Challenges such as imbalanced datasets, adversarial attacks (e.g., URL shortening and homograph attacks), and the need for interpretability in black-box models are discussed. The survey suggests incorporating deep learning and ensemble methods for better accuracy, along with federated learning to preserve user privacy. Additionally, real-time URL scanning and continuous model updates are recommended to counter emerging threats.

The paper "Detection of Malicious Software by Analyzing Distinct Artifacts Using Machine Learning and Deep Learning Algorithms"[3], Ashik et al. examine ransomware and malware detection using machine learning and deep learning models. The study compares algorithms such as Random Forest, SVM, and CNNs in identifying malware based on behavioral and structural artifacts, including API call sequences, file entropy, and registry modifications. The authors note that deep learning models, particularly LSTMs and Transformers, outperform traditional ML methods in detecting zero-day attacks due to their ability to learn complex patterns. However, computational overhead and the need for large labeled datasets remain significant challenges. The paper also discusses the trade-off between detection accuracy and false positives, emphasizing the need for explainable AI techniques in cybersecurity. Future work should focus on developing lightweight deep learning models for real-time deployment, adversarial robustness against evasion techniques, and automated feature extraction to reduce manual engineering efforts.

The survey "A Survey of the Recent Trends in Deep Learning Based Malware Detection"[4] by Priv et al. reviews advancements in deep learning for malware identification. The study discusses techniques like RNNs, LSTMs, and GANs in analyzing malware behavior and evasion tactics, such as code obfuscation and polymorphism. The authors argue that deep learning offers superior detection capabilities compared to signature-based methods, particularly for detecting novel and metamorphic malware. However, they highlight critical issues such as model interpretability, high computational costs, and susceptibility to adversarial attacks (e.g., gradient-based evasion). The survey suggests that future research should explore hybrid models combining deep learning with explainable AI techniques for better transparency. Additionally, unsupervised and self-supervised learning approaches could reduce dependency on labeled datasets, while reinforcement learning might enable adaptive defense mechanisms against evolving threats.

Recent studies have explored various techniques for detecting malicious URLs and software. Maheshwari et al. [1] analyzed URL structures to identify phishing and harmful websites, highlighting the effectiveness of syntax-based detection but noting limitations against obfuscated links. Hoi et al. [2] surveyed machine learning approaches for URL classification, emphasizing the superiority of deep learning models but acknowledging challenges like adversarial evasion and data imbalance. Ashik et al. [3] compared ML and DL algorithms for malware detection, finding that neural networks outperform traditional methods in identifying zero-day threats, though computational costs remain high. Priv et al. [4] reviewed deep learning trends in malware analysis, advocating for hybrid models to enhance interpretability and robustness against evolving cyber threats. Collectively, these studies underscore the need for adaptive, real-time detection systems that balance accuracy, efficiency, and explainability in cybersecurity.

3) Methods and System Implementation

3.1 System Overview

The proposed system detects malicious websites and malware-infected files in real-time using deep learning models. A Chrome extension captures URL and file download requests, sending them to a Flask-based server. The server uses an LSTM model for website classification and a deep neural network for malware detection. Based on the results, the system either allows or blocks the user's request instantly. This approach ensures a safer browsing experience by providing proactive, real-time cybersecurity protection.

3.2 Data Collection and Preprocessing

- **Malicious URLs Dataset**:
  Sourced from Kaggle, containing 651,191 URLs with metadata and threat labels
- **Ransomware Dataset**:
- Containing 62,000+ Windows PE files labeled as benign or ransomware.

Preprocessing Steps:

- Feature extraction (e.g. URL length, HTTPS presence, special characters).
- Normalization and scaling.
- Data splitting inti training and testing sets.

3.3 Machine Learning Models

- **LSTM-based URL Classification Model:**
  Trained to classify URLs as benign, phishing, malware, or defacement. The model uses sequential learning to capture URL patterns and achieved an accuracy of 95.92%.
- **Deep Neural Network for Malware Detection:**
  Designed to detect malware in downloaded files using extracted static features. The model includes multiple dense layers with dropout regularization and achieved 98.48% accuracy for ransomware and malicious files.

3.4 System Architecture

The system initiates with two input streams: URL Requests and Download File Requests, both originating from a Chrome Extension. Each request type is independently processed and analyzed:

- URL requests are evaluated to classify them as safe or malicious.
- Download file requests are inspected to determine if they contain malware.
  Both streams are analyzed by a tested model, and based on the prediction results, actions are taken:
- Safe URLs are allowed.
- Malicious URLs are blocked.
- Safe file downloads are permitted.
- Files containing malware are blocked to ensure system security.

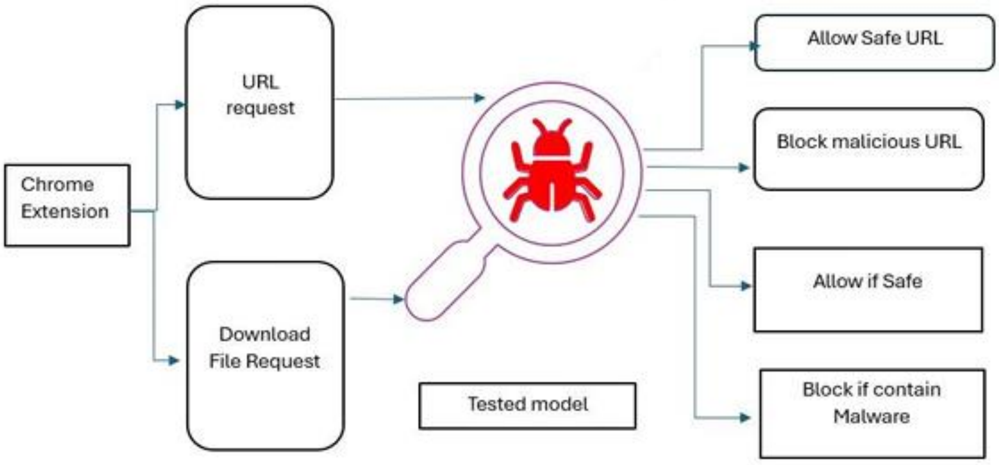

Figure 3.4: System Architecture for Malware and Malicious website detection

4) RESULTS AND DISCUSSION

4.1 Comparative Analysis with Existing Systems

The accuracy between two security tools.

- Comodo Free AV:
  - Accuracy: 78%
  - Lower accuracy among the two options shown.

- URL Threat Detector:
  - Accuracy: 97%
  - Significantly higher accuracy compared to Comodo Free AV.

- Ransomeware Detector:
  - Accuracy: 98.48%
  - Demonstrates superior ability to detect and prevent ransomware threats effectively.

- Existing System (Comodo Free AV):
  - Achieves only 78% accuracy for both URL and ransomware threats, lacking advanced learning capacity.

- Proposed System:
  - LSTM effectively captures sequential URL patterns, resulting in 93% detection accuracy.
  - Deep Neural Networks with dense layers and dropout outperform standard antivirus tools in ransomware identification with 98.48% accuracy.

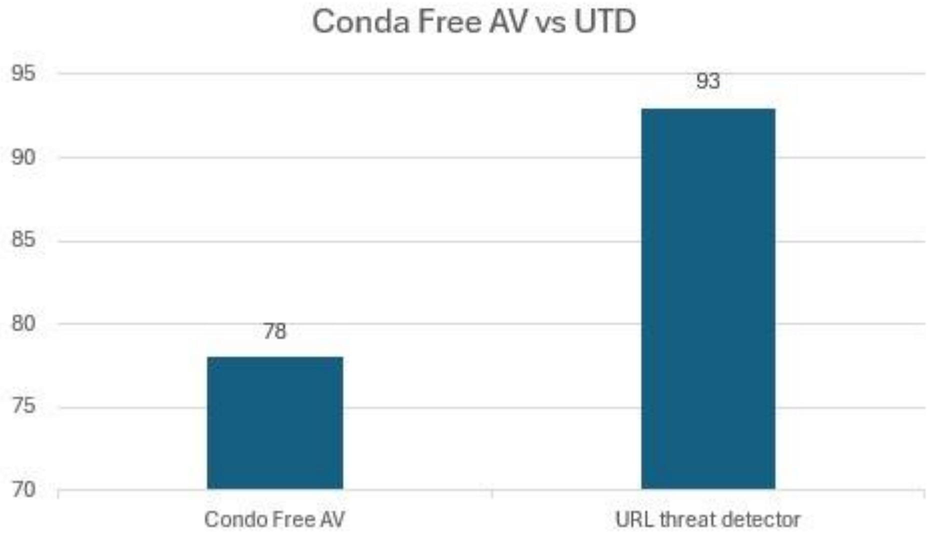

Figure 4.1: Accuracy Comparison between Comodo Free AV and URL Threat Detector

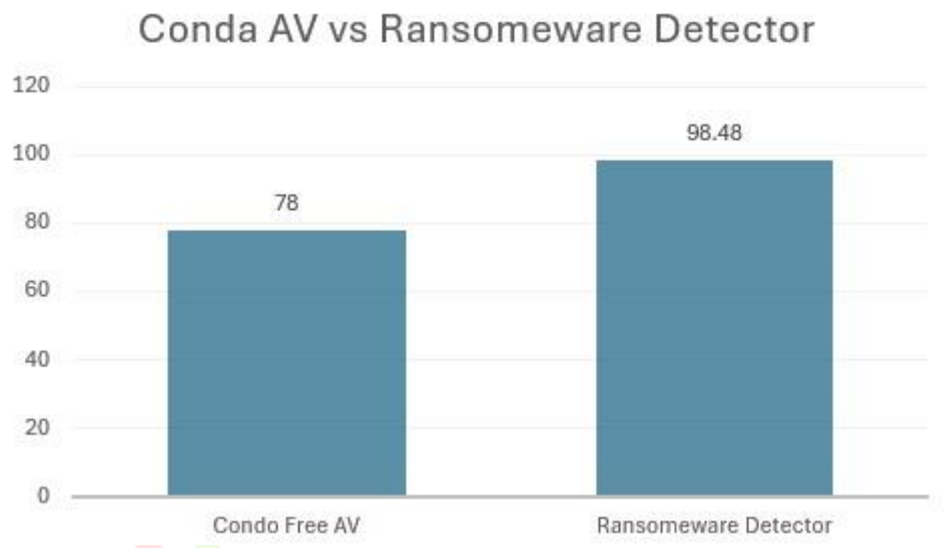

Figure 4.2: Accuracy Comparison between Comodo Free AV and Ransomeware Detector

The bar graph illustrates that both the Ransomware Detector and URL Threat Detector outperform Comodo Free AV in terms of accuracy. While Comodo Free AV achieves 78%, the Ransomware Detector reaches 98.48% and the URL Threat Detector 93%. This indicates that specialized detection tools provide significantly improved protection against evolving cyber threats. Overall, the proposed system demonstrates higher reliability and effectiveness

4.2 Observations and Insights

- The Ransomware Detector significantly outperforms Conda Free AV by achieving 98.48% accuracy compared to 78%, showcasing superior capability in identifying and mitigating ransomware threats.
- Similarly, the URL Threat Detector surpasses traditional AV tools, reaching 93% detection accuracy, effectively filtering malicious URLs before user access.
- The combination of sequential URL analysis (via LSTM) and static feature-based malware detection (via Deep Neural Networks) provides a multi-layered security approach.

- LSTM’s capability to capture subtle sequential patterns in URLs led to significantly higher accuracy in identifying phishing, malware, and defacement threats compared to traditional blacklist methods.
- Deep Neural Networks, trained on static file features, enhanced malware detection by identifying new and obfuscated threats that signature-based methods often miss.
- The integration with a Chrome Extension enables real-time threat detection for both web browsing and file downloads, providing immediate protection and minimizing user risk.
- The system is modular, scalable, and can be deployed across various platforms, including browsers, enterprise environments, and endpoint protection systems for broader cybersecurity coverage.

5) Conclusion and Future Work

5.1 Conclusion:

The Malware and Malicious Website Detection System represents a significant advancement in addressing the growing challenges of cybersecurity threats posed by malicious websites and malware-infected downloads. By leveraging deep learning models for URL classification and file-based malware detection, the system offers a proactive and intelligent solution aimed at safeguarding users from phishing attacks, ransomware, identity theft, and data breaches.

The integration of real-time threat detection through a Chrome extension, combined with a robust backend analysis server, ensures a dynamic and responsive security framework that enhances browsing safety and protects personal information. Through features such as accurate threat classification, automated blocking of harmful activities, and seamless user interaction, the system empowers individuals to navigate the internet more securely. Beyond its technical contributions, the project makes a real-world impact by reducing exposure to cyber risks, promoting safer online practices, and setting a foundation for future innovations in intelligent, data-driven cybersecurity systems.

5.2 Future Scope:

The future scope of this project involves improving the accuracy of threat detection by incorporating more diverse datasets and refining the deep learning models. It could also include extending the system to support other browsers and platforms beyond Chrome. Real-time updates from global threat intelligence sources could further enhance the system's ability to detect emerging cyber threats.

6) REFERENCES

[1] Maheshwari, Shantanu and Janet, B and Kumar, R, “Malicious URL Detection: A Comparative Study”, 1147-1151, 2021

[2] Doyen Sahoo, Chenghao Liu, and Steven C.H. Hoi, “Malicious URL Detection using Machine Learning: A Survey”, 1, 1, 2019

[3] Ashik, M.; Jyothish, A.; Anandaram, S.; Vinod, P.; Mercaldo, F.; Martinelli, F.; Santone, “A. Detection of Malicious Software by Analyzing Distinct Artifacts Using Machine Learning and Deep Learning Algorithms.”, Electronics, 2021

[4] Priv, J. Cybersecurity., “A Survey of the Recent Trends in Deep Learning Based Malware Detection”, 2, 4, 800-829, 2022