“Subterfuge Sentry: Guarding Against Sneaky Malware Tactics”

Ms. Samruddhi Babasaheb Shinde
Dr. Ms. S. P. Pawar
Department of Computer Science Engineering
SVERI’s College of Engineering, Pandharpur, Maharashtra, India.

Abstract- In an era where online security is paramount, this project aims to develop a robust system for detecting the security status of web links in real-time. The proposed solution leverages a browser extension equipped with four distinct machine learning algorithms - Support Vector Machines (SVM), Random Forest, Boosting, and Multilayer Perceptron (MLP). Each algorithm contributes its unique strengths in analyzing various features associated with web links, providing a comprehensive evaluation of their security.

The project begins by collecting a diverse dataset containing labeled instances of secure and insecure links, incorporating a wide range of features such as URL structure, SSL/TLS certificate information, and historical threat intelligence data. The dataset is then used to train and fine-tune each algorithm, optimizing their ability to accurately classify links into secure and insecure categories.

Following the individual training of the algorithms, a novel approach is introduced to enhance overall accuracy - ensemble learning. The ensemble model employs a Voting algorithm that combines the predictions of SVM, Random Forest, Boosting, and MLP. This collaborative decision-making process maximizes the strengths of each algorithm, mitigating potential weaknesses and creating a more resilient and reliable link security detection system.

The browser extension seamlessly integrates into users' web browsers, providing real-time feedback on the security status of links encountered during browsing. Users are alerted to potential security threats, allowing them to make informed decisions when interacting with web content.

The project's success is evaluated through extensive testing and validation using diverse datasets and real-world scenarios. The overall accuracy of the system is measured, demonstrating the effectiveness of the ensemble learning approach in enhancing link security detection. The results of this project contribute to the ongoing efforts to enhance online security and empower users with the tools needed to navigate the digital landscape securely.

***Key Words*:** Link Security, Ensemble Learning, Machine Learning, Support Vector Machines (SVM), Random Forest, Boosting, Multilayer Perceptron (MLP), Voting Algorithm, Browser Extension, Web Security, Online Threat Detection, Real-time Classification, Feature Extraction, Accuracy Assessment, Cybersecurity, Threat Intelligence, Online Safety.

INTRODUCTION

In an era marked by pervasive digital connectivity, the security of online activities is of paramount importance. One significant aspect of this security landscape is the detection of potentially harmful web links that may lead to security threats such as phishing, malware, or other cyber-attacks. This project introduces a novel approach to address this concern by employing a browser extension equipped with advanced machine learning algorithms for real-time link security detection.

Four distinct algorithms - Support Vector Machines (SVM), Random Forest, Boosting, and Multilayer Perceptron (MLP) - are integrated into the system, each contributing its unique strengths to analyze various features associated with web links. These features include aspects such as URL structure, SSL/TLS certificate information, and historical threat intelligence data. By leveraging the strengths of these algorithms, we aim to create a comprehensive and effective link security detection system.

To further enhance the accuracy of our system, we introduce an ensemble learning approach. The ensemble model employs a Voting algorithm that combines the predictions of individual algorithms, thereby creating a robust decision-making mechanism. This collaborative approach mitigates the limitations of individual algorithms and improves the overall accuracy of link security classification.

The project is not only focused on the technical aspects of algorithmic implementation but also emphasizes user accessibility and convenience. A user-friendly browser extension is developed, seamlessly integrating into popular web browsers. This extension provides real-time feedback to users, alerting them to the security status of links encountered during browsing. Empowering users with this information enable them to make informed decisions and navigate the digital landscape securely. Through rigorous testing, validation, and evaluation using diverse datasets and real-world scenarios, we aim to demonstrate the efficacy of our system. The success of this project contributes to the ongoing efforts to fortify online security, offering users a reliable tool to navigate the web with confidence in the face of evolving cyber threats.

LITERATURE SURVEY

1. A Cascade Approach" presents a comprehensive exploration of malware detection methodologies, emphasizing the novelty of their proposed cascade approach. The main findings reveal that the cascade one-sided perceptron (COS-P) algorithm, including its mapped and kernelized versions, exhibits promising accuracy and sensitivity in distinguishing malware from benign files. The study showcases the limitations of conventional signature-based techniques in handling diverse malware behaviors and highlights the need for advanced detection mechanisms. By organizing sources into themes, it becomes evident that the progression of research in this field revolves around enhancing the effectiveness of machine learning-based detection methods, particularly focusing on ensemble techniques and feature engineering to improve classification performance. The key strengths of the paper lie in its systematic comparison of various COS-P adaptations, comprehensive experimentation, and the introduction of algorithmic optimizations to enhance scalability. However, some weaknesses include the absence of real- time testing and an exhaustive analysis of potential false positives. In conclusion, the cascade approach demonstrates its potential as an advanced malware detection solution, showcasing significant advancements over traditional methods and underscoring the value of ensemble learning and algorithmic cascades in cybersecurity.[1]

2. "Malware Detection & Classification using Machine Learning" addresses the crucial issue of malware detection in the current digital landscape. In light of the escalating risk posed by constantly evolving and polymorphic malware, the paper emphasizes the need for effective detection techniques beyond traditional signature-based methods. Recognizing the limitations of conventional tools in tackling dynamic malware behaviors, the authors propose leveraging Machine Learning (ML) techniques for detection. The paper outlines a methodology involving the extraction of behavioral patterns through static or dynamic analysis, followed by the application of diverse ML algorithms to determine whether a given file is malware or not. The study examines both behavioral-based detection methods and the potential of ML algorithms to create a social-based malware recognition and classification model. The authors highlight the significance of ML-driven detection due to the growing volume of novel malware, presenting an urgent need for improved cybersecurity measures. They categorize various types of malwares, such as adware, spyware, viruses, worms, Trojans, rootkits, ransomware, and more, underlining the diverse threat landscape. The paper also discusses malware discovery techniques, dividing them into signature-based and behavior-based approaches. It details the features used for analysis and outlines several machines learning algorithms, including Support Vector Machines (SVM), Naive Bayes, K-Nearest Neighbors (KNN), and Decision Trees, applied in different studies for malware detection and

classification. Experimental results demonstrate the superiority of ML-driven techniques over traditional signature-based approaches, with enhanced accuracy and efficiency in malware detection. Overall, the paper underscores the potential of Machine Learning to revolutionize malware detection and offers valuable insights into various algorithms that have shown promise in this domain.[2]

3. The paper presents a comprehensive study on the effectiveness of deep neural networks, specifically DenseNet, for detecting malware through a visual feature approach. The authors investigate the vulnerability of deep learning models to adversarial attacks, focusing on Gaussian noise and the Fast Gradient Sign Method (FGSM). Using benchmark datasets, the proposed DenseNet model achieves high accuracy and F1-scores for malware detection. The study evaluates the model's resilience to adversarial attacks and showcases its robustness against poisoning and evasion attacks. The research sheds light on the importance of developing malware detection systems that can withstand adversarial attempts, contributing to the advancement of secure computing environments. limitation is the focus solely on visual features while not considering other essential malware detection features, potentially affecting the real-world applicability of the proposed DenseNet approach.[3]

4. Incorporating a hybrid approach, the study employs a Voting Classifier to effectively amalgamate numerous machine learning models for classification through majority voting, thus creating a singular robust classifier. This approach demonstrates manageable execution times, rendering it potentially suitable for extensive malware analysis, particularly in large-scale applications.[4]

5. A static malware detection system that leverages data mining techniques. The study evaluates the effectiveness of SVM, J48, and Naïve Bayes classifiers in detecting malware. Interestingly, results indicate that classifiers based on the DLL name feature exhibit notably poor

detection rates. Furthermore, the Naïve Bayes classifier consistently demonstrates subpar detection accuracy across various scenarios. This study sheds light on the intricate interplay between data mining methods and their application to static malware detection, highlighting the nuanced performance variations among different classifiers and features.[5]

AIM & OBJECTIVES

- Developing Machine Learning Models.
- Feature Extraction and Dataset Creation.
- Ensemble learning integration.
- Browser Extension Development**.**
- Real time link security detection.

MOTIVATION

The motivation behind this project stems from the ever-growing significance of cybersecurity in the digital age. As our lives become increasingly interconnected and reliant on online platforms, the risks associated with malicious activities such as phishing, malware, and cyber-attacks have also escalated. The motivation for developing a robust link security detection system using machine learning and ensemble learning algorithms.

SCOPE

The scope of this project is multifaceted, encompassing various aspects of machine learning, cybersecurity, and user interaction. The project aims to deliver a comprehensive link security detection system with a focus on real-time protection and user empowerment.

PROBLEM DEFINITION

In the digital age, users face an escalating risk of encountering malicious web links that can lead to a range of security threats, including phishing attacks, malware infections, and other cyber-attacks. Traditional methods of link analysis often struggle to keep pace with the evolving tactics employed by cybercriminals, leaving users vulnerable to online threats. The problem addressed by this project is the lack of an efficient and real-time link security detection system that can accurately classify the security status of web links and empower users to make informed decisions during their online activities.

SYSTEM ARCHITECTURE

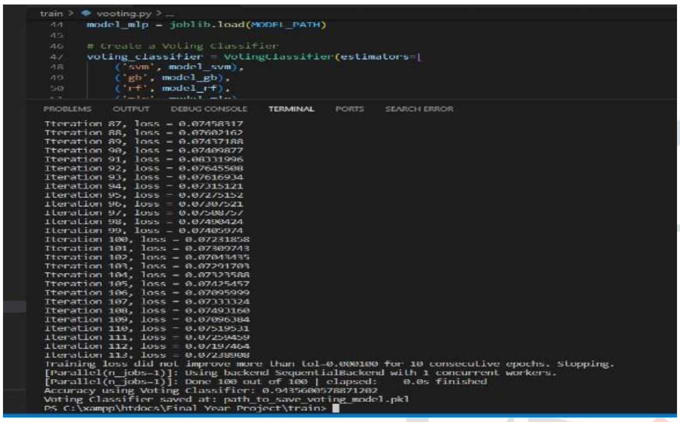
Fig -1: System Architecture Diagram

- The user interacts with the system through a browser extension interface. The extension seamlessly integrates into popular web browsers, ensuring a user-friendly experience.
- This module is responsible for conducting real-time link security assessments. It interfaces with the machine learning models and utilizes ensemble learning techniques to classify links as either secure or insecure.
- The system incorporates four machine learning algorithms: Support Vector Machines (SVM), Random Forest, Boosting, and Multilayer Perceptron (MLP). Each model is trained on a diverse dataset containing labeled instances of secure and insecure links.
- The ensemble learning module combines the predictions of individual machine learning models (SVM, Random Forest, Boosting, MLP) using a Voting algorithm. This collaborative decision-making process aims to improve overall link security detection accuracy.
- The system extracts relevant features from web links, including URL structure, SSL/TLS certificate details, and historical threat intelligence data. These features serve as inputs to the machine learning models, contributing to the accuracy of link security assessments
- The link security detection module communicates in real-time with the browser extension interface, providing instantaneous feedback to the user as they navigate the web. This ensures timely alerts and empowers users to make informed decisions.

- In the case of a potential security threat, the system triggers an alert mechanism within the browser extension interface. Users receive clear and actionable alerts, informing them of the security status of the link they are about to access.

RESULT

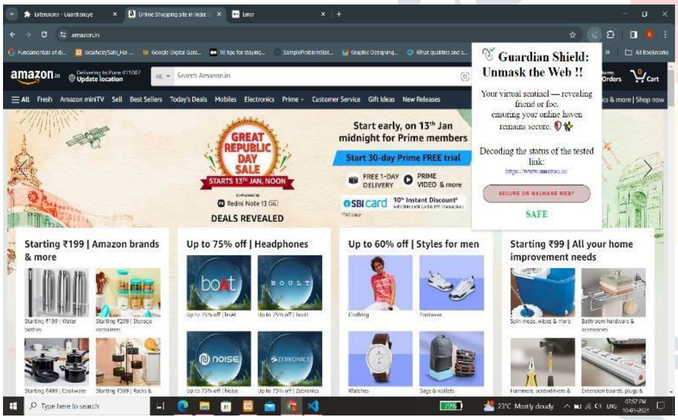
Final Voting Classifier Accuracy:



```
train > vooting.py > ...
44      model_mlp = joblib.load(MODEL_PATH)
45
46      # Create a Voting Classifier
47      voting_classifier = VotingClassifier(estimators=[
48          ('svm', model_svm),
49          ('gb', model_gb),
50          ('rf', model_rf),
```

```
Iteration 87, loss = 0.07458317
Iteration 88, loss = 0.07602162
Iteration 89, loss = 0.07437188
Iteration 90, loss = 0.07409877
Iteration 91, loss = 0.08331996
Iteration 92, loss = 0.07645508
Iteration 93, loss = 0.07616934
Iteration 94, loss = 0.07315121
Iteration 95, loss = 0.07275152
Iteration 96, loss = 0.07307521
Iteration 97, loss = 0.07508757
Iteration 98, loss = 0.07490424
Iteration 99, loss = 0.07405974
Iteration 100, loss = 0.07231858
Iteration 101, loss = 0.07309743
Iteration 102, loss = 0.07043435
Iteration 103, loss = 0.07291703
Iteration 104, loss = 0.07323588
Iteration 105, loss = 0.07425457
Iteration 106, loss = 0.07095999
Iteration 107, loss = 0.07333324
Iteration 108, loss = 0.07493160
Iteration 109, loss = 0.07096384
Iteration 110, loss = 0.07519531
Iteration 111, loss = 0.07259459
Iteration 112, loss = 0.07197464
Iteration 113, loss = 0.07238908
Training loss did not improve more than tol=0.000100 for 10 consecutive epochs. Stopping.
[Parallel(n_jobs=1)]: Using backend SequentialBackend with 1 concurrent workers.
[Parallel(n_jobs=1)]: Done 100 out of 100 | elapsed:    0.0s finished
Accuracy using Voting Classifier: 0.9435600578871202
Voting Classifier saved at: path_to_save_voting_model.pkl
PS C:\xampp\htdocs\Final Year Project\train>
```

Final Output:



Extensions - Guardianeye
Online Shopping site in India
Error
amazon.in
Update location
Search Amazon.in
Orders
Cart
GREAT REPUBLIC DAY SALE
STARTS 13TH JAN, NOON
DEALS REVEALED
Start early, on 13th Jan midnight for Prime members
Start 30-day Prime FREE trial
SBI card
Starting ₹199 | Amazon brands & more
Up to 75% off | Headphones
Up to 60% off | Styles for men
Starting ₹99 | All your home improvement needs
Type here to search
23°C Mostly cloudy

Guardian Shield: Unmask the Web !!

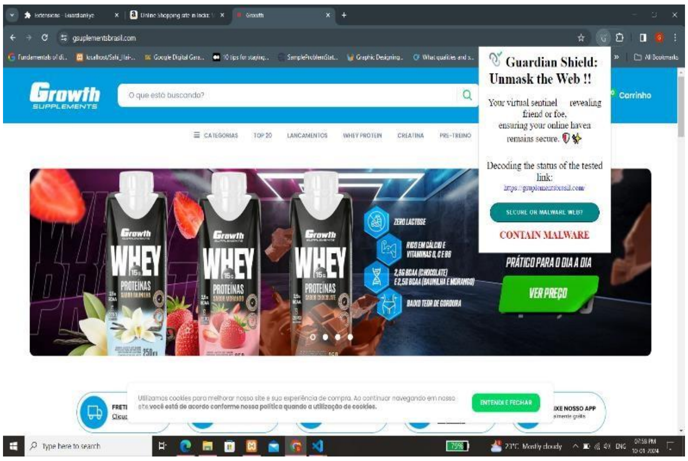
Your virtual sentinel — revealing friend or foe, ensuring your online haven remains secure.

Decoding the status of the tested link:

https://www.amazon.in/

SECURE OR MALWARE WEB?

SAFE



Guardian Shield: Unmask the Web !!

Your virtual sentinel revealing friend or foe, ensuring your online haven remains secure.

Decoding the status of the tested link:

https://gsuplementsbrasil.com/

SECURE OR MALWARE WEB?

CONTAIN MALWARE

CONCLUSION

In conclusion, the development of a link security detection system utilizing ensemble machine learning algorithms and a user-friendly browser extension represents a significant stride towards enhancing online security. The comprehensive approach of this system, combining advanced machine learning techniques with real-time user interaction, addresses the evolving challenges posed by malicious web links.

REFERENCES

[1] N. Milosevic, “History of malware,” 02 2013.

[2] Internet Crime Report, 2021, https://www.ic3.gov/

[3] Mouhammd Alkasassbeh, Mohammad A. Abbadi, Ahmed M. AlBustanji. " LightGBM Algorithm for Malware Detection." Applied Sciences volume 1230 (2022)

[4] 14. Or-Meir, O.; Nissim, N.; Elovici, Y.; Rokach, L. Dynamic malware analysis in the modern era—A state of the art survey. ACM Comput. Surv. 2019, 52, 1–48. [CrossRef] 15. Albulayhi, K.; Abu Al-Haija, Q.; Alsuhibany, S.A.; Jillepalli, A.A.; Ashrafuzzaman, M.; Sheldon, F.T. IoT Intrusion Detection Using Machine Learning with a Novel High Performing Feature Selection Method. Appl. Sci. 2022, 12, 5015.

[5] Document management – portable document format – part 1: Pdf 1.7. Standard, International Organization for Standardization, Geneva, CH, Mar. 2008.

[6]PDF properties and metadata, Adobe Acrobat Accessed 6,Dec 2022

[7] Aslan, Ömer & Samet, Refik. (2020). A Comprehensive Review on Malware Detection Approaches. IEEE Access. 8. 1-1. 10.1109/ACCESS.2019.2963724.

[8] Elingiusti, Michele & Aniello, Leonardo & Querzoni, Leonardo. (2018). PDF-Malware Detection: A Survey and Taxonomy of Current Techniques. 10.1007/978-3-319-73951-9_9.

[9] Albahar, Marwan & Thanoon, Mohammed & Alzilai, Monaj & Alrehily, Alla & Alfaar, Munirah & Al-Ghamdi,

Maimoona & Alassaf, Norah. (2021). Toward Robust Classifiers for PDF Malware Detection. Computers, Materials and Continua. 69. 2181-2202. 10.32604/cmc.2021.018260.

[10] VirusTotal https://virustotal.com/.

[11] Contagio Malware Dump, "External data source," [Online]. Available: http://contagiodump.blogspot.com.au

[12] Falah, Ahmed & Pan, Lei & Huda, Shamsul & Pokhrel, Shiva & Anwar, Adnan. (2021). Improving malicious PDF classifier with feature engineering: A data-driven approach. Future Generation Computer Systems. 115. 314-326. 10.1016/j.future.2020.09.015.

[13] CIC-Evasive-PDFMal2022 Dataset CIC-Evasive-PDFMal2022 | Datasets | Canadian Institute for Cybersecurity | UNB

[14] Abu Al-Haija, Q.; Odeh, A.; Qattous, H. PDF Malware Detection Based on Optimizable Decision Trees. Electronics 2022, 11, 3142.

[15] Chandran, P. & Hema, Rajini & Jeyakarthic, M.. (2022). Invasive weed optimization with stacked long short term memory for PDF malware detection and classification. International journal of health sciences. 4187- 4204. 10.53730/ijhs.v6nS5.9540. Kumar, Akshi. (2018). Machine Learning from Theory to Algorithms: An Overview. Journal of Physics: Conference Series. 1142. 012012. 10.1088/1742-6596/1142/1/012012.