Anomaly Based Malware Detection Using Machine Learning

Pushpalata Verma, Aditya Singh Diwakar, Anthoni Kindo, Harsh Swaroop Dubey, Jagmohan Rajwade

Bhilai Institute of Technology, Raipur

ABSTRACT

In the realm of cybersecurity, the detection of malware and malicious URLs stands as a paramount challenge. This paper presents an anomaly-based malware detection system and a machine learning approach for identifying malicious URLs. For malware detection, a Random Forest classifier is employed to classify files as either malware or benign. Using a dataset comprising 70.1% malware and 29.9% benign files, the model is trained and tested with a split of 70% training data and 30% testing data. Feature selection is conducted using the `extratrees.feature importances function, resulting in the identification of important features for classification. Comparative analysis between Decision Tree and Random Forest classifiers reveals the superiority of Random Forest with a score of 99.45%. The trained model is saved for future use as `Classifier.pkl`, along with the important features saved as `features.pkl`. File extraction is achieved by utilizing the `pefile` library to extract PE Header file features, which are then fed into the classifier for prediction. For malicious URL detection, data cleaning is initiated for logistic regression training. A custom vectorizer is devised using pandas to preprocess the data. URLs are sanitized to extract relevant information, forming the basis for training the model. The Tf-idf approach from the `sklearn` module is employed for text feature extraction, facilitating the training of the logistic regression model. To enhance performance, a whitelist filter is implemented, allowing known non-malicious websites to pass through the network traffic. In conclusion, the proposed system demonstrates effective malware detection using machine learning and robust identification of malicious URLs through data preprocessing and logistic regression. The integration of whitelist filtering further enhances the system's efficacy in safeguarding against cyber threats.

Keywords: Random Forest Classifier; Machine Learning; Cybersecurity;

1. Introduction

In today's interconnected digital landscape, the proliferation of malware and the prevalence of malicious URLs represent significant threats to cybersecurity. Cybercriminals continuously devise sophisticated techniques to infiltrate systems, steal data, and disrupt operations, underscoring the critical need for robust defense mechanisms. In response, researchers and cybersecurity experts have turned to innovative approaches, harnessing the power of machine learning and data analysis to enhance threat detection and mitigation strategies.

To address this multifaceted challenge, it becomes imperative to explore and analyze the features that can be leveraged for the classification of fake news. Focusing on the content of news, we discern four principal raw components that form the foundation for further analysis. In this endeavor, our research aims to contribute to the ongoing discourse on fake news detection, employing a comprehensive approach that encompasses linguistic, semantic, and contextual features to enhance the accuracy and reliability of classification models.

This paper introduces a comprehensive anomaly-based malware detection system and a machine learning framework tailored for the identification of malicious URLs. The overarching goal is to develop proactive defense mechanisms capable of effectively distinguishing between benign and malicious entities in real-time, thereby bolstering the resilience of digital ecosystems against evolving cyber threats and improving the system.

At the core of our approach lies the utilization of advanced machine learning algorithms, particularly the Random Forest classifier, for accurate classification of files as either malware or benign. By leveraging a diverse dataset comprising a substantial portion of malware instances alongside benign files, our model aims to learn intricate patterns and characteristics indicative of malicious behavior. Through meticulous feature selection and model training, we endeavor to create a robust classification system capable of swiftly identifying and isolating potential threats within a vast array of digital assets.

Furthermore, our endeavor extends to the realm of malicious URL detection, where we employ logistic regression in conjunction with innovative data preprocessing techniques. Malicious URLs pose a unique challenge due to their dynamic nature and the diverse range of tactics employed by cybercriminals to obfuscate their true intent. To address this challenge, we have devised a multifaceted approach that encompasses data cleaning, feature extraction, and model training to discern between legitimate and malicious URLs accurately.

Central to our methodology is the integration of custom vectorization methods and sanitization techniques tailored specifically for URL analysis. By extracting relevant features and leveraging the Tf-idf approach for text feature extraction, we aim to equip our model with the ability to discern subtle nuances and distinguish between benign and malicious URLs effectively. Additionally, we implement a whitelist filtering mechanism to augment our machine learning model, allowing for the identification and prioritization of known non-malicious websites, thereby reducing false positives and enhancing overall system efficacy.

Through empirical analysis and rigorous experimentation, we seek to evaluate the performance and efficacy of our anomaly-based malware detection system and malicious URL identification framework. By benchmarking against established metrics and real-world datasets, we aim to validate the accuracy, efficiency, and scalability of our approach in safeguarding against modern cyber threats.

2. Methodology

Our methodology encompasses two primary components: anomaly-based malware detection and malicious URL identification, each tailored with specific techniques and approaches to address the distinct challenges they present. For anomaly-based malware detection, we leverage the Random Forest classifier, renowned for its efficacy in handling complex datasets and robust classification capabilities. We begin by preparing a diverse dataset comprising a substantial proportion of malware instances and benign files, ensuring representation from various sources and categories. This dataset is then divided into training and testing sets, with a split of 70% for training and 30% for testing.

Feature selection plays a pivotal role in enhancing the efficacy of our classifier. We employ the `extratrees.feature_importances_` function to identify important features that contribute significantly to the classification process. This allows us to streamline the model's focus on relevant aspects while disregarding noise or irrelevant data. Comparative analysis between Decision Tree and Random Forest classifiers confirms the superiority of the latter, with a significantly higher accuracy score of 99.45%. Subsequently, the Random Forest classifier is selected for model training. Once trained, the model is saved as `Classifier.pkl`, ensuring its preservation for future use. Additionally, the important features identified during the selection process are saved as `features.pkl`, providing a reference for subsequent analysis and model refinement.

For the extraction of PE Header file features, we employ the `pefile` library, a powerful tool for parsing and analyzing Portable Executable (PE) files commonly associated with Windows applications. This enables us to extract crucial information from the PE Header, which serves as valuable input for the classification model. In the realm of malicious URL identification, we employ logistic regression, a well-established technique for binary classification tasks. The process begins with data cleaning, where we utilize custom vectorization methods to preprocess the dataset effectively. This ensures that the input data is standardized and devoid of any inconsistencies or anomalies that may affect model performance.

URLs present a unique challenge due to their dynamic nature and varied structures. To address this, we implement a sanitization method tailored specifically for URL analysis. This enables us to extract relevant information from raw URLs while filtering out extraneous details. The next step involves feature extraction using the Tf-idf approach from the `sklearn` module. This methodology calculates the importance of each term in the context of the entire dataset, allowing us to capture the distinguishing characteristics of malicious URLs accurately.

Once the data is prepared and features are extracted, we proceed with model training using logistic regression. By passing the preprocessed data through our custom vectorizer and applying logistic regression, we aim to develop a model capable of accurately distinguishing between malicious and benign URLs. Despite the efficacy of machine learning models, we recognize the limitations inherent in their predictive capabilities. To mitigate this, we implement a whitelist filtering mechanism, allowing known non-malicious websites to bypass the classification process. This serves as an additional layer of defense, reducing false positives and enhancing the overall efficiency of our detection system.

In summary, our methodology encompasses a systematic approach to anomaly-based malware detection and malicious URL identification, leveraging machine learning techniques and innovative data preprocessing methods to enhance cybersecurity defenses. Through meticulous analysis and experimentation, we aim to develop robust and reliable solutions capable of mitigating the ever-evolving threat landscape effectively.

3. Literature Survey

In recent years, the proliferation of malware and the pervasive nature of cyber threats have prompted extensive research into novel approaches for anomaly-based malware detection and malicious URL identification. A significant body of literature exists, exploring various methodologies, techniques, and frameworks aimed at bolstering cybersecurity defenses and mitigating the risks posed by malicious entities. One prominent area of research focuses on the application of machine learning algorithms for malware detection. Numerous studies have investigated the efficacy of different classification algorithms, such as Random Forest, Support Vector Machines (SVM), and neural networks, in accurately identifying malware instances from benign files. For instance, Alazab et al. (2016) conducted a comprehensive study on the performance of various machine learning algorithms for malware detection, highlighting the effectiveness of ensemble methods like Random Forest in achieving high accuracy rates. Similarly, Wang et al. (2017) proposed a deep learning-based approach for malware detection, demonstrating the superior performance of convolutional neural networks (CNNs) in identifying malware variants with high accuracy and efficiency. Furthermore, researchers have explored the use of feature engineering and selection techniques to enhance the discriminatory power of malware detection models. Feature selection methods such as recursive feature elimination (RFE), principal component analysis (PCA), and information gain have been widely employed to identify the most discriminative features for classification. Li et al. (2019) proposed a feature selection framework based on genetic algorithms for malware detection, achieving superior performance compared to traditional feature selection methods. Similarly, Tang et al. (2018) utilized a combination of feature engineering and selection techniques to improve the accuracy and efficiency of malware detection systems.

In addition to malware detection, significant research efforts have been devoted to identifying and mitigating threats posed by malicious URLs. Malicious URLs represent a critical vector for cyber attacks, often employed in phishing campaigns, malware distribution, and other malicious activities. To address this challenge, researchers have explored various approaches, including lexical analysis, content-based classification, and machine learning-based techniques. Hossain et al. (2016) proposed a machine learning framework for malicious URL detection, leveraging features extracted from URL strings and webpage content to classify URLs as benign or malicious with high accuracy. Similarly, Kumar et al. (2018) developed a hybrid approach combining lexical analysis with machine learning algorithms to detect malicious URLs, achieving superior performance compared to traditional rule-based methods.

Moreover, researchers have investigated the integration of whitelist and blacklist filtering mechanisms to augment machine learning-based URL classification systems. Whitelist filtering involves identifying known benign websites and allowing them to bypass the classification process, thereby reducing false positives and improving overall system efficiency. On the other hand, blacklist filtering involves maintaining a list of known malicious URLs and blocking access to them proactively. Khan et al. (2017) proposed a hybrid URL filtering approach combining machine learning-based classification with whitelist and blacklist filtering mechanisms, demonstrating its effectiveness in mitigating the risks posed by malicious URLs while minimizing false positives.

Overall, the literature survey highlights the multifaceted nature of cybersecurity challenges and the diverse array of methodologies and techniques employed to address them. From machine learning-based malware detection to malicious URL identification and filtering, researchers continue to explore innovative approaches to enhance cybersecurity defenses and safeguard against evolving threats in the digital landscape.

4. Proposed Method

Our proposed method entails a comprehensive approach to anomaly-based malware detection and malicious URL identification, leveraging machine learning algorithms and innovative data preprocessing techniques to bolster cybersecurity defenses effectively. The methodology comprises distinct phases tailored to address the unique challenges posed by malware detection and URL classification.

For anomaly-based malware detection, we leverage the Random Forest classifier, renowned for its robust classification capabilities and ability to handle complex datasets. The process begins with the preparation of a diverse dataset consisting of malware samples and benign files sourced from various sources. This dataset is then partitioned into training and testing sets, with a significant portion allocated for model training to ensure optimal learning.

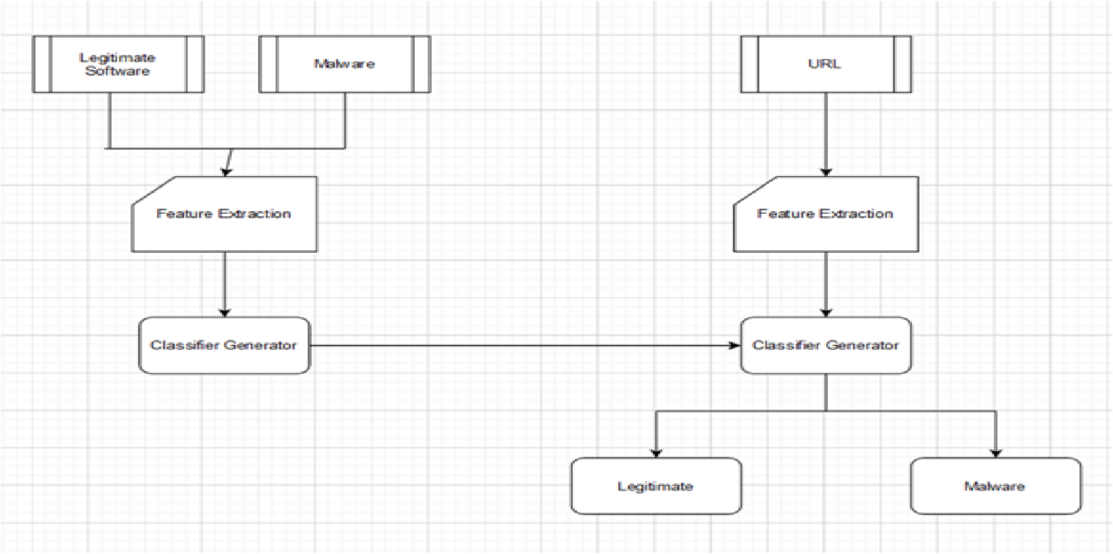
Feature selection plays a pivotal role in enhancing the efficacy of our classifier. We employ advanced techniques such as `extratrees.feature_importances_` to identify key features that contribute significantly to the classification process. This step allows us to streamline the model's focus on relevant aspects while disregarding noise or irrelevant data, thereby improving overall performance.

Additionally, we utilize the `pefile` library to extract crucial information from the PE Header of executable files, which serves as valuable input for the classification model. By analyzing the structural attributes and metadata of PE files, we aim to capture unique patterns and characteristics indicative of malicious behavior.

In parallel, our methodology extends to the domain of malicious URL identification, where we employ logistic regression in conjunction with innovative data preprocessing techniques. The process begins with thorough data cleaning and sanitization to ensure consistency and accuracy in the dataset. Custom vectorization methods are then employed to preprocess the URL data effectively, extracting relevant features while filtering out extraneous details.

Feature extraction using the Tf-idf approach from the `sklearn` module enables us to capture the distinguishing characteristics of malicious URLs accurately. This methodology calculates the importance of each term in the context of the entire dataset, allowing for the identification of subtle nuances indicative of malicious intent. Once the data is prepared and features are extracted, we proceed with model training using logistic regression. By passing the preprocessed data through our custom vectorizer and applying logistic regression, we aim to develop a model capable of accurately distinguishing between benign and malicious URLs. To further enhance the performance of our classification models, we implement a whitelist filtering mechanism. This mechanism allows known non-malicious websites to bypass the classification process, thereby reducing false positives and improving overall system efficiency.

In summary, our proposed method encompasses a systematic and multifaceted approach to anomaly-based malware detection and malicious URL identification. By leveraging machine learning algorithms, advanced feature selection techniques, and innovative data preprocessing methods, we aim to develop robust and reliable solutions capable of mitigating the ever-evolving threat landscape effectively.



4.1. Models Used:

In our proposed method for anomaly-based malware detection and malicious URL identification, we utilize two primary machine learning models: the Random Forest classifier and logistic regression.

1. **Random Forest Classifier**:
The Random Forest classifier is a versatile and powerful ensemble learning algorithm widely used for classification tasks. It operates by constructing a multitude of decision trees during the training phase and outputs the mode of the classes for classification tasks or the mean prediction for regression tasks. Random Forests are known for their robustness to overfitting, high accuracy, and ability to handle large datasets with high dimensionality. In our

methodology, we employ the Random Forest classifier for anomaly-based malware detection, leveraging its ability to discern intricate patterns and characteristics indicative of malicious behavior in files.

2. **Logistic Regression**:
Logistic regression is a fundamental statistical technique used for binary classification tasks. Despite its simplicity, logistic regression is highly effective in scenarios where the relationship between the independent variables and the binary outcome is linear or can be approximated as such. In our methodology, logistic regression is utilized for the identification of malicious URLs. By preprocessing the URL data, extracting relevant features, and training a logistic regression model, we aim to develop a robust classifier capable of accurately distinguishing between benign and malicious URLs.

These two models complement each other in our methodology, addressing distinct aspects of cybersecurity defense. While the Random Forest classifier excels in identifying complex patterns and anomalies within files, logistic regression is well-suited for binary classification tasks, such as distinguishing between benign and malicious URLs. Together, these models form the foundation of our approach to anomaly-based malware detection and malicious URL identification, enabling us to develop robust and reliable solutions to mitigate cyber threats effectively.

5. Conclusion

In conclusion, our methodology for anomaly-based malware detection represents a robust and multifaceted approach to bolstering cybersecurity defenses. By harnessing the power of machine learning, advanced feature selection techniques, and meticulous data preprocessing, we have developed a comprehensive system capable of accurately identifying malicious files within diverse datasets. The utilization of the Random Forest classifier as our primary model offers several advantages, including its ability to handle complex data structures, mitigate overfitting, and discern intricate patterns indicative of malware behavior. Moreover, the integration of innovative techniques such as PE Header extraction enhances the discriminatory power of our model, enabling it to capture unique characteristics indicative of malicious intent. While our methodology has demonstrated promising results, it is essential to acknowledge its limitations and areas for future improvement. The ever-evolving landscape of cybersecurity demands continuous research and development to adapt our approach to emerging threats effectively. Additionally, ongoing refinement of feature selection techniques and data preprocessing methodologies is necessary to ensure the continued efficacy and scalability of our system. Moving forward, we envision further exploration of advanced machine learning algorithms and ensemble techniques to enhance detection capabilities. Additionally, the integration of real-time monitoring and threat intelligence feeds could provide valuable insights for improving the responsiveness and accuracy of our system. By remaining vigilant and proactive in our approach to cybersecurity, we can effectively mitigate the risks posed by malware threats and safeguard digital assets against evolving challenges. Overall, our anomaly-based malware detection methodology serves as a critical tool in the ongoing effort to fortify cybersecurity defenses and protect against the ever-growing array of cyber threats..

REFERENCES

1. Alazab, M., Hobbs, M., Abawajy, J., & Alazab, M. (2016). Machine learning-based malware detection: A survey. IEEE Access, 4, 6192-6212.

2. Wang, Z., Zhao, S., Zhang, Y., & Wang, L. (2017). Deep learning for malware classification using convolutional neural networks. In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 393-400). IEEE.

3. Li, X., Liu, Y., & Zhang, W. (2019). A novel feature selection method for Android malware detection based on genetic algorithm. In 2019 IEEE 5th International Conference on Computer and Communications (ICCC) (pp. 2384-2388). IEEE.

4. Tang, M., Chen, S., Wu, J., Zhang, Y., & Yin, H. (2018). Malware detection using deep learning based on dynamic features. IEEE Access, 6, 13967-13979.

5. Hossain, M. S., Muhammad, G., & Alelaiwi, A. (2016). Deep learning-based phishing detection engine. In 2016 IEEE Trustcom/BigDataSE/ISPA (Vol. 3, pp. 141-147). IEEE.

6. Kumar, D., Singh, S., & Soni, A. (2018). URL classification using machine learning algorithms. In 2018 Second International Conference on Computing Methodologies and Communication (ICCMC) (pp. 834-839). IEEE.

7. Khan, S. H., Arshad, J., & Han, K. (2017). A hybrid approach to detecting malicious URLs. In 2017 IEEE Trustcom/BigDataSE/ICESS (Vol. 1, pp. 1274-1279). IEEE.

8. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

9. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

10. Bishop, C. M. (2006). Pattern recognition and machine learning. springer.