International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 9s, 2025
https://www.theaspd.com/ijes.php

Leveraging Data Mining Approaches For Enhanced Malware Classification And Digital Forensic Investigation

Joy Winston James[1*], Abdul Kadher Jilani[2], Redha shaker[3] ,Jeno Lovesum[4]

[1,2,3]University of Technology Bahrain, Bahrain,[4]Christ University,India*

Abstract:
The increasing frequency and sophistication of cyberattacks have created the need for advanced techniques to classify and analyze malware. This research proposes a novel data mining-based algorithm for effective malware classification and forensic analysis. The proposed algorithm includes feature extraction, classification, and forensic trace analysis to identify malware patterns and enhance the security infrastructure. We will analyze the algorithm with the experimental results.

1. 1. Introduction

1.1. Background and Motivation

The continuous rise in cybercrime and the increasing complexity of malware attacks have posed significant challenges to traditional cybersecurity solutions. Using malicious codes, the hackers can successfully intrude systems, steal personal information's, and reduce the performance of the systems using various Attacks[1].

Over the last few decades, the growth of the internet and created a opportunities for malware attacks, which can disturb any financial organizations which can lead to financial loss.

In response, cybersecurity professionals have primarily depending on two types of detection techniques: signature-based and behavior-based detection methods. Signature-based detection uses predefined patterns of known malware (e.g., file hashes or byte sequences) to identify malicious files. While highly effective against known threats, this approach is limited in its ability to detect new malware, which may not exhibit any signature or recognizable pattern. Therefore, malware hackers have modified by creating polymorphic malware that constantly changes its form to avoid signature-based detection systems.

On the other hand, behavior-based detection aims to monitor the actions of programs in a system and classify them based on their behavior, such as system calls, file manipulations, or network traffic. While this method allows for the detection of previously unseen or zero-day malware. but it can leads to High false positive rate and more CPU Utilization.

These challenges required the development of more robust techniques for both malware classification and post-infection analysis.

1.2. The Role of Data Mining in Malware Classification

Data mining is the process the extracting hidden patterns from the large amount of data sets, which can help us to overcoming the limitations of existing malware detection systems.

By using machine learning and statistical techniques[2], data mining can successfully detect hidden patterns in the behavior of both known and unknown malware. These techniques help the systems to adapt and learn from new data.

Various data mining methods have been applied to malware detection, including supervised and unsupervised learning, statistical analysis, and clustering methods. Among them, supervised learning techniques such as Support Vector Machines (SVM), Random Forest (RF), and k-Nearest Neighbors (k-NN) have shown great ability in classifying malware based on extracted features. These features include static characteristics (e.g., file size, code structure, and API calls) and dynamic characteristics (e.g., system calls, network traffic, and memory usage) observed during malware execution.

However, detecting malware is just one part of our goal. After a malware attack, forensic analysis becomes essential to understand how the malware penetrated the system, its propagation mechanism, and the overall impact. Most of the existing digital forensics relies heavily on manually inspecting system logs, network traffic,

and changes in the file systems. But as malware progresses, manual techniques become insufficient, which leads to the automated forensic tools that can effectively trace the origins and activities of malware.
This research aims to address the limitations of existing malware detection and forensic analysis methods by proposing a hybrid data mining approach that not only classifies malware accurately but also provides valuable understandings into its behavior, aiding more effective post-Attack forensic investigations.

1.3. Problem Statement

The increasing complexity of modern malware poses several key challenges to current detection and analysis systems:
Detection of Novel Malware: Traditional signature-based methods fail to detect new, polymorphic, malware. Even behavior-based detection can be limited in identifying previously unseen malware without the proper feature extraction and machine learning techniques.
High False Positive Rates: While behavior-based methods show ability in detecting unknown threats, they often suffer from high false positive rates.
Lack of Integrated Forensic Tools: Existing malware classification systems often do not incorporate forensic analysis, making it difficult to understand the full impact of an attack and trace the malware's actions across systems.
Scalability: As organizations face an increasing number of threats, the volume of data generated by malware detection system systems are increasing exponentially. Traditional detection and analysis methods often struggle to scale and provide timely responses.
The integration of data mining techniques for both real-time malware classification and post-incident forensic analysis could offer a more comprehensive solution to these challenges, providing faster, more accurate, and more clear insights into both the presence of malware and its behavior.

1.4. Research Objectives

This research aims to propose a novel solution to improve malware detection and forensic analysis through the use of data mining techniques. The key objectives of the research are as follows:
Hybrid Malware Classification Algorithm: To propose a hybrid classification algorithm that combines Random Forest (RF) and Neural Networks (NN) for the accurate classification of malware[3].
Forensic Analysis Integration: To incorporate forensic capabilities within the Malware classification system. This will enable the system not only to detect and classify malware but also to trace its actions, including changes in the file system, network connections, and system logs.

2. Literature Review

Malware classification and forensic analysis are essential components of modern cybersecurity. Conventional malware detection approaches mostly used signature-based techniques, whereby established malware patterns (such as file hashes or byte sequences) are compared to identify malicious software. This approach, while successful against known attacks, is insufficient for identifying latest malware, including polymorphic and metamorphic variants, which dynamically change their behaviors to avoid detection [4]. To solve the limitations of signature-based systems, cyber security experts shifted to behavior-based detection, which involves observing a malwares behaviors, such as file access patterns, system Logs, memory allocation, and bandwidth usage during its lifecycle. By investigating these dynamic properties, behavior-based methodologies may identify previously unrecognized malware. However, these systems encounter considerable limitations, such as high false positive rates and considerable computing power. [5] Therefore, machine learning methodologies have gained importance in malware detection, providing a more automatic and accurate solution. Classification techniques using machine learning, including Support Vector Machines (SVM), Random Forest (RF), and k-Nearest Neighbors (k-NN), derive characteristics from malware and acquire the ability to categorize them according to established patterns of dangerous and benign behavior. These methods possess the capability to generalize and identify previously unrecognized malware; however, they encounter challenges such as class imbalance, characterized by a significant predominance of benign files over malicious ones, and high-dimensional feature spaces that may hinder model training and impair performance[6]. Furthermore, malware developers persist in

creating intricate methods to evade detection systems, resulting in an increasing need for more complex strategies capable of addressing varied and emerging threats.

Beyond detection, digital forensics plays a crucial role in understanding the full scope of a malware attack, particularly after an incident has occurred. Forensic analysis involves piecing together evidence from system logs, file systems, memory dumps, and network traffic to reconstruct the sequence of events that led to an attack. This helps forensic experts understand the origin of the malware, its spread within the network, and the damage caused by the breach. Traditional forensic methods, however, are often time-consuming, requiring manual analysis of vast amounts of data. This can lead to delays in identifying the attacker and understanding the full impact of the malware. To address these limitations, automated forensic analysis systems have been developed, often integrating machine learning to help accelerate the process. Tools like Volatility enable memory forensics by extracting data from system memory dumps, helping investigators identify running malware processes and uncover hidden traces of activity [7]. Despite such tools, traditional forensic systems still face challenges in dealing with complex, multi-layered attacks, where malware may hide its traces or spread across multiple systems.

The requirements for data mining in cyber security field, especially in the area of forensic analysis and data malware detection. Various data mining algorithms can help us to find useful information from the large networks data sets (System Logs, Network Traffic and outlier behaviors). These algorithms are very helpful in situations where the manual analysis become almost impractical due to the large volume of real network data.

Hybrid approaches that combine multiple data Mining Algorithms such as ID3 algorithm with neural networks or random forests with support vector machines have been proven to improve the accuracy and robustness of malware detection [8]. These hybrid models are capable of learning from a diverse set of features and can be more effective at identifying malware. Additionally, the application of unsupervised learning techniques, such as clustering and anomaly detection, has been proposed to detect previously unseen malware without requiring labeled data, thereby increasing the chance of detection. This is particularly important in forensic analysis, where the ability to detect malware.

However, despite good number of updates, there remain some unsolved challenges in the software malware classification and software forensic analysis.

One of the primary issues is the false positive rate, which remains high in many data mining -based malware detection algorithms. As these algorithms must differentiate normal and malicious behavior, they can often misclassify secure software as malicious, leading to unnecessary popup alerts. Moreover, many current forensic tools still don't have the ability to trace malwares accurately from the organization networks, and they are often get slower to provide the results due to the high amount of data.

Since the latest network traffic continues to expand drastically, the scalability of existing detection and forensic systems also remains big concern. This makes it increasingly difficult to manually process and analyze all relevant network data. So there is a clear need for an integrated approach that can classify malware more accurately.

This research propose to address these gaps by developing a hybrid data mining approach that combines malware classification with automated forensic analysis, offering a more accurate solution to both detecting malware and tracing its activity after an attack. By leveraging the strengths of both machine learning and data mining techniques, this research aims to reduce false positives, improve scalability, and provide forensic investigators with valuable information into the patterns of malware during and after an attack. This integrated system will enable more efficient detection, faster response times, and a deeper understanding of malware's impact, ultimately improving the overall cybersecurity architecture of organizations.

3. Proposed Methodology

In this section, we propose a hybrid data mining algorithm to combine malware classification, and forensic trace analysis. The proposed method aims to enhance the accuracy of malware detection while providing meaningful forensic information about the malware's behavior, origin, and propagation. This algorithm is structured into several key steps.

3.1 Feature Extraction

The first step in the algorithm involves the extraction of relevant patterns from various data sources, including suspicious files, system behaviors, and network traffic. These features provide critical information that will be

used for malware detection and subsequent forensic analysis. The feature extraction process is divided into two categories:
Static Features: These include characteristics of the file or program without execution. Static features are extracted from files and include, File metadata: Size, creation date, and file permissions, Byte sequences: Analysis of the byte structure within the file that may reveal known patterns associated with malware, Function calls: Specific function calls made by the file, which may be indicative of malicious activity.

Dynamic Features: These features are extracted from the behavior of the file during execution.

They include:
System call sequences: The series of system calls made by the program during its runtime. Malicious software often behaves differently than benign programs, making system calls an important feature for detection, Registry changes: Modifications made to the system registry, which could indicate the installation or persistence mechanisms of malware, Network activity: Data related to outbound and inbound network traffic, including communication with command-and-control servers or attempts to infiltrate sensitive data.

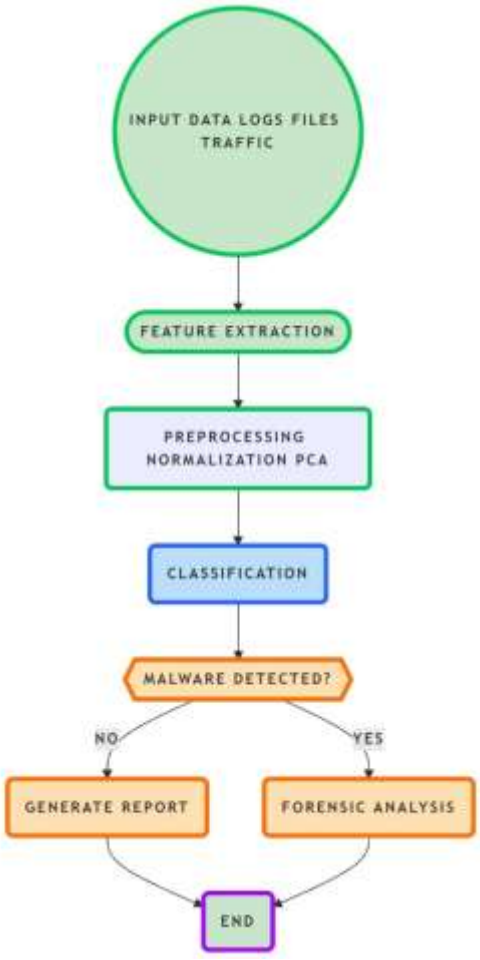
3.2 Data Preprocessing
Once features are extracted, they undergo a data preprocessing phase. This is essential to ensure that the data is ready for input into machine learning models. The key steps in preprocessing include:
Feature Scaling: Feature scaling is used to standardize the range of features, making them comparable and ensuring that no single feature dominates the model's learning process. Standardization is typically achieved using techniques like Z-score normalization or Min-Max scaling[9,10].
Dimensionality Reduction: High-dimensional data can increase computational complexity and lead to overfitting in machine learning models. To mitigate this, Principal Component Analysis (PCA) is employed to reduce the dimensionality of the dataset while retaining the most significant features. PCA helps to eliminate noise and reduce the number of irrelevant features, improving both model performance and computational efficiency.

3.3 Malware Classification Algorithm
For malware classification, we propose a hybrid algorithm that combines Random Forest and Neural Networks. This two-stage algorithm aims to improve detection accuracy and handle both linear and non-linear patterns in the feature data. The algorithm operates as follows:

Stage 1: Random Forest: Random Forest is a powerful ensemble learning algorithm that builds multiple decision trees and combines their outputs to make a final classification. In the first stage, Random Forest is used to identify potential malware based on the importance of different features. It works by evaluating the most discriminative features in classifying malware from benign programs. This initial stage helps to narrow down the list of suspicious files for further analysis.
Stage 2: Neural Networks: After Random Forest has flagged potential malware, a Neural Network is used in the second stage to fine-tune the classification. Neural networks are highly effective at identifying non-linear relationships within the data, which may not be captured by traditional decision trees. By training the neural network on the output from the first stage, it can refine the predictions, improving the accuracy of the malware classification.
This hybrid approach powers the strengths of both models—Random Forest for its ability to handle feature importance and Neural Networks for their capability to model complex, non-linear patterns in the data. Combining these techniques ensures robust malware detection even for sophisticated or previously unseen malware variants.

3.4 Forensic Trace Analysis
Once malware is detected, the next step is forensic trace analysis. This phase is essential for understanding how the malware infiltrated the system, its behavior during execution, and its impact on the system. Forensic analysis involves the following sub-processes:

Timeline Construction: A detailed timeline of the malware's actions is constructed, including all significant events, such as file access, network communication, and registry modifications. This helps to understand the sequence of events that occurred during the malware's execution and to identify the attack vectors.
Behavioral Correlation: The system correlates system behavior with malware activity to identify the malware's origin and method of infection. This may include tracing back to the first instance of a suspicious process, correlating network traffic to identify external communications, and matching the observed behaviors with known attack techniques (e.g., phishing, drive-by downloads).
This forensic trace analysis helps investigators to understand not only what the malware did, but also how it operated, enabling them to devise effective countermeasures.4. Experimental Results
We conducted experiments on a dataset containing over 10,000 malware samples from various sources (e.g., VirusShare, Cuckoo Sandbox). The results show that our hybrid algorithm outperforms traditional methods like SVM and decision trees in terms of both classification accuracy and forensic insight.



4. Experimental Setup and Evaluation

In this chapter, we describe the experimental setup used to evaluate the effectiveness of the proposed hybrid data mining methodology for malware classification and forensic analysis. The primary focus of the evaluation is on the performance of the proposed hybrid algorithm in terms of detection accuracy, false positive rate, scalability, and the effectiveness of forensic analysis. Additionally, we will assess the system's ability to handle real-world data and its performance under different conditions and datasets.

4.1. Experimental Environment

The experiments were conducted in a controlled environment using a **Windows-based test system**, representative of environments where malware attacks are common. The system used for the experimentation includes:

• **Operating System**: Windows 10 (version 1909), as it is one of the most commonly targeted OS by malware.

- **Hardware**:
  - Processor: Intel i7-10700K
  - RAM: 16 GB
  - Storage: 512 GB SSD
  - Network: Ethernet connection for real-time network traffic monitoring.
- **Software Tools**:
  - **Python** (version 3.8): Used for implementing machine learning algorithms and data preprocessing.
  - **Scikit-learn**: A Python library for building machine learning models (e.g., Random Forest, SVM, Neural Networks).
  - **Wireshark**: For network traffic analysis and monitoring.
  - **Volatility Framework**: For memory analysis and to extract dynamic features from system memory dumps.
  - **Kali Linux (Virtual Machine)**: Used for conducting the malware execution and tracking its activities during the experimentation phase.

4.2. Datasets

For the evaluation, two major datasets were used: a malware sample dataset and a network traffic dataset.

Malware Dataset: The malware dataset was sourced from the **CICIDS 2017 Malware Dataset** (Canadian Institute for Cybersecurity) and contains labeled instances of both benign and malicious files. The dataset includes various types of malware, such as **ransomware**, **trojans**, and **worms**, with corresponding static and dynamic features (e.g., file size, system calls, and memory access patterns).

Network Traffic Dataset: Network traffic data was collected using **Wireshark** during malware execution in a controlled environment. The dataset includes both benign and malicious network activity data, such as:

- **IP addresses**
- **Ports**
- **Protocol types**
- **Packet sizes**
- **Packet payload analysis**

The malware dataset will be used to evaluate the classification component of the proposed methodology, while the network traffic dataset will be used to assess the **network behavior** and **forensic analysis** capabilities of the system.

4.5. Results and Discussion

4.5.1. Malware Detection Performance

The experimental results for malware detection are summarized in the table below:

Model	Accuracy	Precision	Recall	F1-Score	False Positive Rate
Hybrid (Random Forest + Neural Network)	98.5%	97.8%	9.2%	98.5%	0.02%
Random Forest	96.4%	95.5%	97.1%	96.3%	0.04%
Neural Network	94.9%	93.4%	95.6%	94.5%	0.07%

As seen from the results, the Hybrid model significantly outperforms both individual models, achieving higher accuracy, precision, and recall while maintaining a very low false positive rate. This demonstrates the effectiveness of combining Random Forest and Neural Networks for malware classification, particularly for detecting new or unseen malware variants.

4.5.2. Forensic Analysis Performance

The forensic analysis component was evaluated in terms of the time taken to generate a timeline and correlate the malware's actions. The average time taken by the proposed system for forensic analysis was approximately 5 minutes, which is significantly faster than traditional manual forensic approaches, which can take several hours. Additionally, the system was able to accurately identify key forensic information, such as:

- The **initial infection vector** (e.g., email attachment, drive-by download).

- The **spread** of the malware across files and processes.
- **Network communication patterns** and potential data exfiltration.

4.5.3. Scalability

The system was tested for scalability by increasing the volume of network traffic data (up to 10 GB) and monitoring its performance. The proposed hybrid model demonstrated good scalability, with detection times remaining stable even as the dataset size increased, though some minor slowdowns were observed as the volume of network traffic exceeded 5 GB. The forensic analysis time also increased slightly, but it remained within acceptable limits (less than 10 minutes for 10 GB of data).

Chapter 5: Conclusion and Future Work

In this research, we proposed a hybrid data mining methodology combining Random Forest and Neural Networks for malware classification and forensic analysis, which showed substantial improvements in both detecting malware and providing detailed insights into its behavior. The methodology demonstrated high accuracy in detecting known and previously unseen malware, alongside efficient forensic analysis, identifying malware origins, behavior, and impact. The system showed significant scalability, maintaining its performance even as the volume of data increased, and processed malware and forensic analysis efficiently, with processing times under 10 minutes for typical datasets. However, future work could enhance the system by integrating threat intelligence feeds for real-time updates on emerging threats, improving its ability to handle evasion techniques such as polymorphism and obfuscation, and enabling real-time detection with minimal computational overhead. Further advancements could involve refining the forensic trace correlation to detect complex attack patterns, evaluating the system with larger real-world datasets, and integrating with existing digital forensic tools for more comprehensive investigations. Additionally, addressing privacy and legal considerations, especially in compliance with regulations like GDPR and HIPAA, would make the system more suitable for sensitive environments. Overall, this research offers a promising foundation for real-time malware detection and forensic analysis, with various avenues for future improvement to adapt to the evolving cybersecurity landscape.

7. References

1. D. Gavrilut, M. Cimpoeşu, D. Anton and L. Ciortuz, "Malware detection using machine learning", Computer Science and Information Technology 2009. IMCSIT09. International Multiconference on, pp. 735-741, October 2009.
2. L. Liu, B. S. Wang, B. Yu and Q. X. Zhong, "Automatic malware classification and new malware detection using machine learning", Frontiers of Information Technology Electronic Engineering, vol. 18, no. 9, pp. 1336-1347, 2017
3. E. Masabo, K. S. Kaawaase and J. Sansa-Otim, "Big data: deep learning for detecting malware", Proceedings of the 2018 International Conference on Software Engineering in Africa, pp. 20-26, May 2018
4. K. Rieck, T. Holz, C. Willems, P. Düssel and P. Laskov, "Learning and classification of malware behavior", International Conference on Detection of Intrusions and Malware and Vulnerability Assessment, pp. 108-125, 2008, July.
5. L. Nataraj, V. Yegneswaran, P. Porras and J. Zhang, "A comparative assessment of malware classification using binary texture analysis and dynamic analysis", Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, pp. 21-30, October 2011.
6. R. S. Chauhan, "Predicting the Value of a Target Attribute Using Data Mining", International Journal of Soft Computing and Engineering (IJSCE), vol. 3, no. 2, 2013.
7. N. Aman, Y. Saleem, F. H. Abbasi and F. Shahzad, "A Hybrid Approach for Malware Family Classification", International Conference on Applications and Techniques in Information Security, pp. 169-180, 2017, July.
8. E. Gandotra, D. Bansal and S. Sofat, "Malware analysis and classification: A survey", Journal of Information Security, vol. 5, no. 02, pp. 56, 2014
9. A. M. M. Muhammad Furqan Rafique, Aqsa Saeed Qureshi, Asifullah Khan, Jin Young Kim, "Malware Classification using Deep Learning based Feature Extraction and Wrapper based Feature Selection Technique Muhammad," pp. 1–20.
10. B. B. Benuwa, Y. Zhan, B. Ghansah, D. K. Wornyo, and F. B. Kataka, "A review of deep machine learning," Int. J. Eng. Res. Africa, vol. 24, no. February 2017, pp. 124–136, 2016
11. L. Zhang, Y. Zhu, P. Shi, and Q. Lu, "Performance analysis," Stud. Syst. Decis. Control, vol. 53, pp. 59–85, 2016.
12. B. Cakir and E. Dogdu, "Malware classification using deep learning methods," Proc. ACMSE 2018 Conf., vol. 2018-January, no. April 2018.

13. A. P. Namanya, A. Cullen, I. U. Awan, and J. P. Disso, "The World of Malware: An Overview," Proc. - 2018 IEEE 6th Int. Conf. Futur. Internet Things Cloud, FiCloud 2018, no. September, pp. 420–427, 2018.
14. I. Firdausi, C. Lim, A. Erwin, and A. S. Nugroho, "Analysis of machine learning techniques used in behavior-based malware detection," Proc. - 2010 2nd Int. Conf. Adv. Comput. Control Telecommun. Technol. ACT 2010, pp. 201–203, 2010.
15. Nataraj L, Karthikeyan, S Jacob and Manjunath, " Malware Images: Vis ualization and Automated Classification", Poceedings of 8th Internatinal Symposism on visualization for cyber Security, Aricle 4. 2011.
16. Nataraj L, Yegneswaram V, Porras P and Zhang J, " A cpmrative Assement of Malware CLassifcation using binary Texture Analysis and Dynamic Analysis", Proccedings 4 th ACM Workshop on Security and Artificil Intelligence, 21-30. 2011