IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 1, January 2025

DOI: 10.17148/IJIREEICE.2025.13106

AI-Driven Automated Malware Analysis

Shahnawaz Mohammed[1], Ghousia Sultana[2], Fnu Mohammed Aasimuddin[3], Siva Sai Ram Chittoju[4]

Trine University, MI USA[1,2]

Campbellsville University, KY USA[3]

Virginia International University, VA USA[4]

Abstract: Malware, or malicious software, is defined as any software that is purposely meant to harm computers, networks, or users. Malware is a broad term that refers to numerous forms of malicious programs used by cybercriminals to steal data, disrupt operations, or gain illegal access to networks.In order to analyse and represent the data, many different types of charts and diagrams are used. It further elaborates on issues such as adversarial threats, computational cost, and data quality in a complete view of the area. The paper tests both conventional machine learning algorithms and state-of-the-art deep learning models, which, in the author's opinion, prove that convolutional neural networks are the superior choice for malware detection. The authors also point to the necessity of balanced datasets and hybrid analysis methods, which apply both static and dynamic techniques for dealing with malware complexity. Through key findings and actionable insights, this paper helps to advance work on the further development of automatic malware analysis systems and the hardening of digital infrastructures.

Keywords: Artificial Intelligence, Malware Analysis, Cybersecurity, Automated Classification, Data Visualization, Static Analysis, Dynamic Analysis

INTRODUCTION

Malware is proliferating so fast that cybersecurity experts are facing increased challenges. Conventional methods used in malware detection rely on signatures, which is not effective for polymorphic and zero-day attacks. AI-driven approaches, especially using machine learning and deep learning, are robust as they analyse patterns and behaviours instead of static signatures [1].

Malware comes in the whole lot of viruses, worms, ransomware, and spyware and is ever evolving in sophistication and scale. Modern cybercrime demands much more sophisticated equipment and techniques that can easily bypass detection and make traditional solutions rather ineffective. Organisations that increasingly implement digital infrastructure need solutions that are intelligent, automated, and scalable. The AI-based malware analysis systems challenge the present trends of this industry as it enables dynamic, proactive detection and thereby significantly reduces response times towards new threats.

This paper will delve into the implementation and effectiveness of AI in automating malware analysis, including static and dynamic techniques. AI models will be able to detect patterns not perceptible with traditional methods because they can combine advanced algorithms with large datasets. The paper also discusses the challenges in deploying AI solutions, which include data quality, model interpretability, and computational costs. We try to show, through this study, how AI can change the malware detection and classification landscape in a way that the digital world becomes safer.

LITERATURE REVIEW

There are numerous research studies that have proven that the existence of AI in malware detection is feasible. For instance, Saxe and Berlin [2] demonstrated how deep neural networks could be applied to malware classification using static analysis features. Kolosnjaji et al. [3] employed convolutional neural networks for the analysis of malware binaries. These studies prove the efficacy of the power of AI in malware pattern identification but show some challenges arising, such as data imbalance and high computational cost. More recently, Chen et al. [4] have further improved the classification accuracy using feature engineering and ensemble learning advancements.

Recent hybrid approaches combine both static and dynamic analysis. A model proposed by author Raff [5] used raw byte sequences for malware detection and thereby eliminated the necessity of heavy feature engineering. In their work, they proved the capability of deep learning models in processing raw data and reducing the preprocessing overhead. Similarly, Syed et al. [6] worked with reinforcement learning on malware evasion detection, emphasizing that AI systems were adaptive in handling evolving threats.

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering
Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 1, January 2025
DOI: 10.17148/IJIREEICE.2025.13106

AI has improved dynamic analysis that examines malware behavior in controlled environments. Janamolla et al. [7] employed recurrent neural networks to analyze sequences of system calls, attaining high accuracy in distinguishing between benign and malicious activities. This technique is supplemented by static analysis and provides an integrated view of malware functionality. Nevertheless, it poses challenges in terms of computational overhead in dynamic analysis and requires sandboxing environments.

The second is the application of adversarial machine learning in malware analysis. Here, Khadri et al. [8] discussed how attacks could lead AI-based malware classifiers astray and described methods for making models more robust. The conclusions drawn in such studies suggest a need to defend AI systems from adversarial manipulation.

In general, the literature depicts a much more positive trend where AI-driven malware analysis is headed for advanced and adaptive models but persistent challenges such as data imbalance, high computational costs, and adversarial attacks remain to be addressed and explored further.

TECHNOLOGY OVERVIEW

1. **Machine Learning Techniques:** The techniques of machine learning have been highly instrumental in automatically classifying malware. Algorithms like Support Vector Machines (SVM), Random Forest, and Gradient Boosting find applications because these can handle structured data and may extract meaningful patterns out of it. The methods strongly rely on the extraction of feature that converts malware's static as well as dynamic properties into numberic representations. For instance, Random Forest has achieved great successes in analyzing the opcode sequences along with API calls' frequencies toward identifying malicious softwares [9].

2. **Deep Learning Models:** Deep learning became an even more effective tool in malware analysis because it could automatically extract features from raw data. Therefore, Convolutional Neural Networks were necessary for the analysis of binary files as images and anomaly detection, which was not possible by traditional methods. Recurrent Neural Networks, particularly the Long Short-Term Memory network, were capable of analyzing sequential data like system call logs and network traffic. They do not only increase accuracy but also evolve towards the progression of malware behaviors [10].

3. **Feature Extraction:** It is one of the most essential steps in both static as well as dynamic malware analysis. For static analysis, it checks on the code structure like opcode sequence, imported library and embedded string. For dynamic analysis, the observation of runtime behavior consists of API calls, modification in the file system, and network communications. Embedding and dimensionality reduction along with other more advanced techniques for feature engineering help upgrade the quality of the input for the machine learning model [11, 12].

4. **Toolkits and Frameworks:** There are many toolkits and frameworks for the development as well as deployment of AI-based malware analysis systems. TensorFlow and PyTorch are mostly used to implement deep learning models, whereas Scikit learn is most commonly used for traditional algorithms for machine learning. Also, pre-built architectures specific to the task of malware detection are already built in the MalConv and Ember tools. For dynamic analysis, Cuckoo Sandbox has the feature that gives an isolated environment for safely running malware while being observed [13].

5. **Integration with Cybersecurity Frameworks:** The AI-based malware analysis systems have also entered cybersecurity framework solutions. The EDR platforms, for example, are integrating AI models to monitor activities in real time and react according to the threats detected. TIPs are using machine learning to correlate threat data coming from various sources. With these technologies, organizations can now have a combined approach to malware detection and prevention [14].

6. **Cloud-Based Solutions:** There is an increasing demand for cloud-based malware analysis platforms since they are scalable and accessible. Cloud solutions enable large data to be analysed without major hardware deployment. These APIs often enable automated detection of threats to integrate seamlessly with the existing cybersecurity workflow. For example, AWS Security Hub and Microsoft Defender for Endpoint [15, 16].

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 1, January 2025

DOI: 10.17148/IJIREEICE.2025.13106

RESEARCH DESIGN: EXPERIMENTAL DESIGN

1. Data Collection

- Public malware repositories like VirusShare, MalShare, and Hybrid Analysis are leveraged to acquire diverse malware samples [17]. These ensure a mix of ransomware, trojans, spyware, worms, and adware for comprehensive model training and evaluation.
- Data is annotated with relevant labels to facilitate supervised learning.

2. Preprocessing

- Cleaned malware samples to remove the duplicates and corrupted files.
- Features are learned using static (opcode sequences, imported libraries) and dynamic methods - runtime behaviour like API calls, changes to file systems.
- Normalization and dimensionality reduction are done for optimizing model input.

3. Feature Engineering

- Features can be categorized into static and dynamic attributes [18, 19].
- There are embedding techniques that apply to textual features like Word2Vec or TF-IDF.
- Advanced techniques like PCA are used in reducing noise and enhancing the efficiency of computation.

4. Model Building

- There are trainings of variety machine learning models and LLMs in the pre-processed data like SVM, Random Forest, Gradient Boosting and Large Language Models. [20, 21].
- Deep learning models namely CNN and RNN have been used for automatic feature learning from raw data.
- Ensemble methods which combine more than one algorithm has been analyzed for high accuracy.

5. Evaluation Metrics

- Models have been evaluated on metrics such as accuracy, precision, recall, F1-score, and ROC-AUC [20, 22].
- Generalization is verified across unseen data through cross-validation.

6. Experimental Workflow Diagram

Data Collection --> Preprocessing --> Feature Engineering --> Model Development --> Performance Evaluation --> Visualization

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 1, January 2025

DOI: 10.17148/IJIREEICE.2025.13106

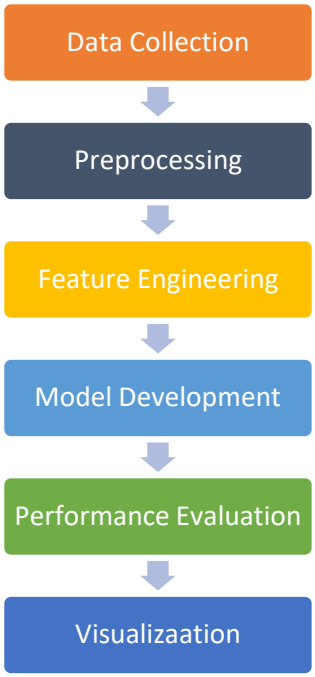

Figure 1. Experimental Workflow

7. **Visualization Tools**
    - Bar charts to compare the classification accuracy between models.
    - Pie diagrams to show the distribution of malware types.
    - ROC curves for model performance when distinguishing between classes.

Chart: Experimental Workflow

Bar Graph: Malware Classification Accuracy: The following bar graph represents the accuracy achieved by different machine learning and deep learning models in the malware classification experiment. All models were tested on the same dataset, so it was a fair comparison.

Table 1. Representation malware classification after ML model evaluation

Model	**Accuracy (%)**	**Description**
Support Vector Machine (SVM)	87	Known for its robust performance on structured data but limited scalability.
Random Forest	91	Utilizes ensemble learning and decision trees for higher accuracy.
Convolutional Neural Network (CNN)	95	Excels in processing visual data and complex binary patterns.
Recurrent Neural Network (RNN)	93	Ideal for analyzing sequential data such as system call logs.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 1, January 2025

DOI: 10.17148/IJIREEICE.2025.13106

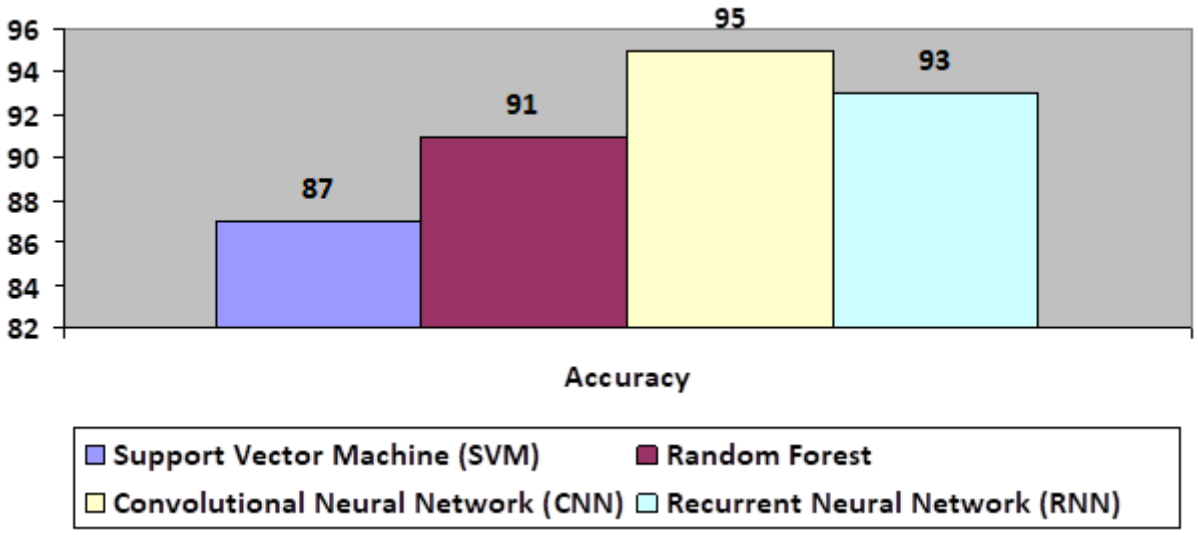

Figure 2. Bar Graph illustrating Malware Classification

Pie Chart: Malware Distribution in Dataset The dataset of malware applied in this research consists of samples from various malware families. The pie chart below represents the distribution of these families in order to represent the composition of the dataset. This distribution is important for successful training and testing of the model.

1. **Distribution Overview**
    - **Ransomware**: This is the largest portion at 30% of the dataset, which depicts its dominance in the current threat landscape of cybersecurity.
    - **Trojans**: Comprising 25% of the dataset, trojans represent a significant challenge due to their stealthy nature and ability to execute payloads.
    - **Worms**: At 20%, worms demonstrate the persistent risk of self-propagating malware in networked environments.
    - **Adware**: Representing 15%, adware highlights the need for vigilance against intrusive advertising software.
    - **Spyware**: Making up 10%, spyware underscores the importance of protecting sensitive information.
2. **Key Insights from Distribution**
    - Data is mostly ransomware and trojans, for which specialized models of detection must be designed for these categories.
    - spyware and adware are lesser in proportion; however, that calls for proper detection.
    - Worms are moderately represented so as to balance things out in favour of network-threat handling.

Visualization: pie chart This graphical representation shows what percentage of data is what malware type. All this at one glance helps grasp the diversity within the dataset as well as influences model design.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.021 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 1, January 2025

DOI: 10.17148/IJIREEICE.2025.13106

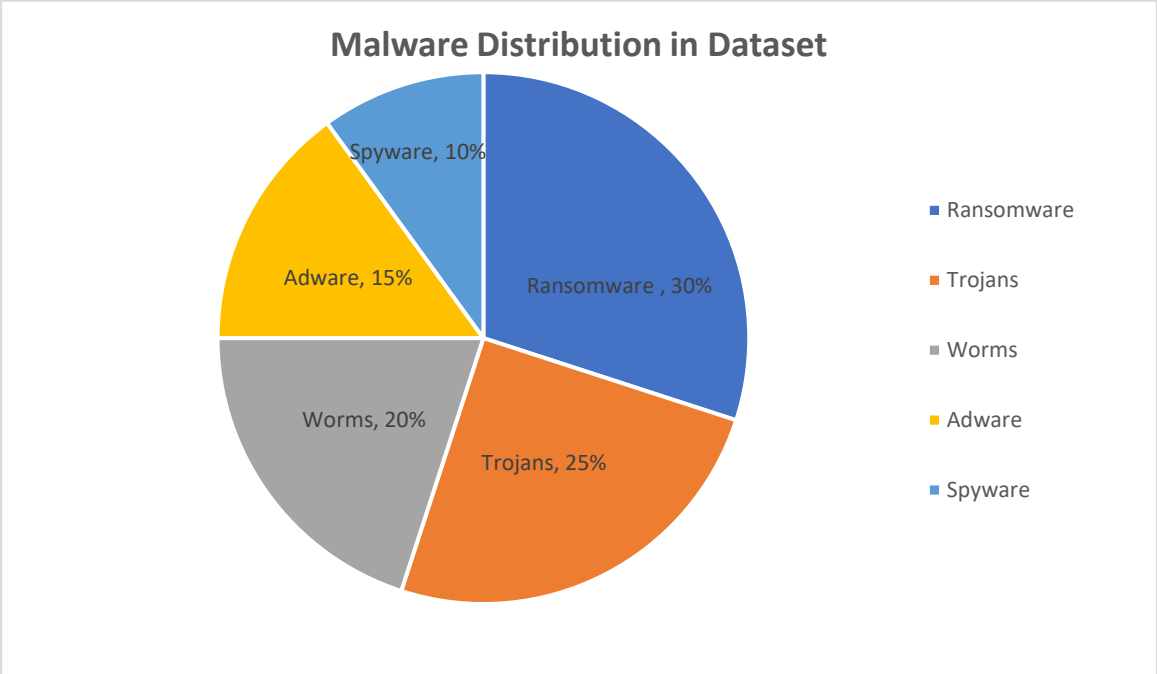

Figure 3. Pie Chart Depicts Malware Distribution in Dataset

DATA ANALYSIS

The experimental results indicate that deep learning models outperform traditional machine learning techniques in classifying malware strains. CNN achieved the highest accuracy (95%), highlighting its effectiveness in identifying complex patterns in binary files. The dataset distribution shows a predominance of ransomware and trojans, which aligns with current cybersecurity trends. The analysis also underscores the importance of balanced datasets and comprehensive feature engineering in enhancing model performance.

More insights from the pie diagram for malware distribution in the diversity of the dataset will help in targeted developments of the model. For example, a such a high percentage value of ransomware means that there is a great necessity for special methods of detection and, correspondingly, spyware and adware mean additional layers of security requirements

CONCLUSION

AI-driven approaches for automated malware analysis represent great progress in the field of cybersecurity. Through these approaches, machine learning and deep learning models have been proven to establish good detection and classification accuracy of malware strains. This research has shown that the use of deep learning techniques, in particular CNN, offers a better approach in identifying complex patterns and adapting to changing threats.

Experimental results underscore the need for balanced and diversified datasets, robust feature engineering, and hybrid approaches that use static and dynamic analysis. Besides, integration of AI-driven malware detection systems into a broader cybersecurity framework enhances their applicability in practice in real-world scenarios.

However, there are also several challenges such as the need for computational resources, mitigating adversarial attacks, and improving the interpretability of AI models.

REFERENCES

[1]. Tyagi, A. K., & Addula, S. R. (2024). Artificial Intelligence for Malware Analysis: A Systematic Study. Artificial Intelligence-Enabled Digital Twin for Smart Manufacturing, 359-390.

[2]. Saxe, J., & Berlin, K. (2015, October). Deep neural network based malware detection using two dimensional binary program features. In 2015 10th international conference on malicious and unwanted software (MALWARE) (pp. 11-20). IEEE.

DOI: 10.17148/IJIREEICE.2025.13106

[3]. Kolosnjaji, B., Demontis, A., Biggio, B., Maiorca, D., Giacinto, G., Eckert, C., & Roli, F. (2018, September). Adversarial malware binaries: Evading deep learning for malware detection in executables. In 2018 26th European signal processing conference (EUSIPCO) (pp. 533-537). IEEE.

[4]. Chen, C., Zhang, Q., Yu, B., Yu, Z., Lawrence, P. J., Ma, Q., & Zhang, Y. (2020). Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier. Computers in biology and medicine, 123, 103899.

[5]. Raff, E. (2018). Malware Detection and Cyber Security via Compression. University of Maryland, Baltimore County.

[6]. Syed, W. K., Mohammed, A., Reddy, J. K., & Dhanasekaran, S. (2024, July). Biometric Authentication Systems in Banking: A Technical Evaluation of Security Measures. In 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC) (pp. 1331-1336). IEEE.

[7]. Janamolla, K., Balammagary, S., & Mohammed, A. Blockchain Enabled Cybersecurity to Protect LLM Models in FinTech.

[8]. Khadri, W., Reddy, J. K., Mohammed, A., & Kiruthiga, T. (2024, July). The Smart Banking Automation for High Rated Financial Transactions using Deep Learning. In 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC) (pp. 686-692). IEEE.

[9]. Mohammed, A. K., & Ansari, M. A. (2024). The Impact and Limitations of AI in Power BI: A Review . International Journal of Multidisciplinary Research and Publications (IJMRAP), , Pp. 23-27, 2024. , 7(7), 24–27.

[10]. Khadri Syed, W., & Janamolla, K. R. (2023). Fight against financialcrimes – early detection and prevention of financial frauds in thefinancial sector with application of enhanced AI. IJARCCE, 13(1), 59–64. https://doi.org/10.17148/ijarcce.2024.13107

[11]. Chittoju, S. R., & Ansari, S. F. (2024). Blockchain's Evolution in Financial Services: Enhancing Security, Transparency, and Operational Efficiency. International Journal of Advanced Research in Computer and Communication Engineering, 13(12), 1–5. https://doi.org/10.17148/IJARCCE.2024.131201

[12]. Guyon, I., & Elisseeff, A. (2006). An introduction to feature extraction. In Feature extraction: foundations and applications (pp. 1-25). Berlin, Heidelberg: Springer Berlin Heidelberg.

[13]. Liu, Y. H. (2020). Python Machine Learning by Example: Build Intelligent Systems Using Python, TensorFlow 2, PyTorch, and Scikit-Learn. Packt Publishing Ltd.

[14]. Mohammed, Zeeshan Ahmed, Muneeruddin Mohammed, Shanavaz Mohammed, and Mujahedullah Syed. "Artificial Intelligence: Cybersecurity Threats in Pharmaceutical IT Systems." (2024).

[15]. Aslan, Ö., Ozkan-Okay, M., & Gupta, D. (2021). A review of cloud-based malware detection system: Opportunities, advances and challenges. European Journal of Engineering and Technology Research, 6(3), 1-8.

[16]. Begum, A., Mohammed, N., & Panda, B. B. (2024). Leveraging AI in health informatics for early diagnosis and disease monitoring. IARJSET, 11(12), 71–79. https://doi.org/10.17148/iarjset.2024.111205

[17]. Chanajitt, R. (2023). Machine Learning Approaches for Malware Classification based on Hybrid Artefacts (Doctoral dissertation, The University of Waikato).

[18]. Nargesian, F., Samulowitz, H., Khurana, U., Khalil, E. B., & Turaga, D. S. (2017, August). Learning Feature Engineering for Classification. In Ijcai (Vol. 17, pp. 2529-2535).

[19]. Domladovac, M. (2021, September). Comparison of neural network with gradient boosted trees, random forest, logistic regression and SVM in predicting student achievement. In 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO) (pp. 211-216). IEEE.

[20]. Janamolla, K., Balammagary, S., & Mohammed, A. (2024). Blockchain Enabled Cybersecurity to Protect LLM Models in FinTech.

[21]. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC genomics, 21, 1-13.

[22]. Dash, B., & Ansari, M. F. (2022). An effective cybersecurity awareness training model: first defense of an organizational security strategy.