Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Malware Classification using Dynamic Analysis and Machine Learning

Harshit Mishra
Reseach Scholar, M. Tech CSE, Shri Ramswaroop Memorial University, Lucknow, UP, India

Neelesh Mishra
Assistant Professor, Shri Ramswaroop Memorial University, Lucknow, UP, India

Abstract - Malware is one of the biggest threats in the modern era of digital world, and since the attacks are becoming more advanced, making them difficult to be detected. The goal of this study is to detect and categorise malware using machine learning and dynamic analysis methods. This method runs malware samples in a controlled environment called sandbox, to see how they actually behave in real time, including system calls, file operations, and their network actions. Machine learning models then uses those behavioural patterns as the input features. The malware or virus is then categorized using different types of techniques that includes Naive Bayes, Decision Tree, Random Forest, and Support Vector Machine. The result shows the best accuracy of machine learning models among them, particularly Random Forest and SVM, in detecting and categorising malware. The study also shows that machine learning and dynamic analysis methods gives more accurate and efficient result upon working together then the word separately. Upon considering all the things, this study provides a more reliable and improved approach of detecting modern cyber threats.

Keywords: Malware classification, Dynamic analysis, Sandbox environment, API calls, System calls, Machine learning, Deep learning, Cybersecurity.

1. INTRODUCTION

Computers and the internet are used in almost every area of life in today's modern and digital world, which includes communication, commerce, finance, and education. Event they are benefitting us but on the other hand they increase the risk of cyberattacks. One of the most common and dangerous risk is malware. It is a category of malicious software that is aimed to compromise systems, steal information (including person details or files), or access devices without authorization. Since the malware has become more advanced and complex by using modern tools and methods, it becomes more difficult to detect them and prevent cyberattacks.

Previously used techniques like the signature-based methods, for the malware detection, are of no longer efficient. These methods are not able to detect new malwares or the viruses that are not detected till now, which are commonly known as the zero-day attacks, as because these techniques are based on already known patterns of malware and cannot detect the new patterns. Therefore, new techniques are required that are more flexible and complex that they can detect new malwares patterns that are seen for the first time.

One of the methods that can understand a program’s behaviour by running that program in a controlled environment preventing them to cause any harm if they are intended to do so, is the Dynamic analysis technique. This technique monitors the behaviour of the program while executing instead of just checking the code. This process involves keeping the track on its behaviour like its system calls, network activities and file modifications etc. By monitoring these behaviours of the program, it becomes simpler to determine that whether the program is safe or not.

Machine learning had become a very useful and efficient tool for cybersecurity in the last few years. It is able to detect patterns in the data without being programmed explicitly and make choices on their own. When Machine learning is combined with Dynamic analysis technique then it can help in detecting and classifying malware based on its behaviour. This approach makes the process of detection and categorizing, faster, accurate and efficient, also this technique is capable of detecting new malwares and threats that are unknown.

The main aim of this study is to examine that how machine learning and dynamic analysis can categorize the malware more effectively. In order to categorize different samples of malware, this study monitor the behaviour of those malwares and uses different machine learning methods. This research focuses on improving the accuracy and efficiency of malware detection system.

Importance of advanced methods to prevent modern cyber threats are highlighted in this study. It also shows that how combining the machine learning and dynamic analysis can help in creating a stronger and more reliable security systems.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

2. LITERATURE REVIEW

Many people have researched on malware detection and classification over past many years. They basically focus on machine learning and dynamic analysis. This section covers the important research in this area and also explains their methods, conclusions and limitations clearly.

Previous studies shows that attackers use different strategies like obfuscation and packaging to hide malicious code so that it can be prevented from being detected. This is the reason that makes the traditional approaches ineffective. Survey research identifies three primary categories of malware analysis methods which include memory-based analysis and dynamic analysis and static analysis. Because it examines the actual behaviour of malware while it is run in a controlled environment, dynamic analysis is seen to be more beneficial among these. [1]

Numerous research emphasize how crucial machine learning is for detecting malware. While machine learning can identify unknown threats by learning patterns from existing data, traditional signature-based approaches are unable to detect new infections. According to a review by Bala Krishna et al., machine learning models identify malware more successfully than conventional methods by using behavioural patterns like API calls and system activity. However, the quality of the data and the features that are chosen have a significant impact on these models. [2]

Several academics, in particular, have focused on integrating machine learning and dynamic analysis. For example, one study used sandbox-based dynamic analysis to gather malware behavior data and applied different machine learning methods which include Random Forest and Naive Bayes and Decision Trees. The results showed a very high classification accuracy with performance that is nearly perfect sometimes. But the study also showed some drawbacks which includes dataset with unclear sizes and the risk of overfitting. This could affect the model performance on new data. [3]

Another important study focused on supervised machine learning methods for examine and detecting dynamic malware. The research showed that the malware can be categorised effectively by using classifiers like Random Forest and Support Vector Machines which primarily focuses on runtime behaviour of the malware. The study also found that dynamic traits provided more detailed insights on the virus's activity but they also required more time and processing power. [4]

Deep learning has been applied for detecting and classifying malware in a recent study. One of the studies that was proposed using a hybrid model which mixes Long Short-Term Memory and Convolutional Neural Networks together to boost up the performance of classification. This approach when compared to usual approaches, improves the detection of malware by capturing spatial and sequential patterns in the malware data. But however, the deep learning models are complex, requires a large dataset and are also hard to understand. [5]

Many of the studies have compared the Static and dynamic analytic methods. One study found that the dynamic analysis methods remain important. It reveals some hidden behaviour which is difficult to be detected or found by static methods. Sometimes the static features work better than the dynamic features. The study showed that combining these both strategies might improve the overall performance of the model. While the benefit is not always large, it suggests that the hybrid models need a careful design which can help. [6]

Researcher found that classifying the malware into categories are very hard. Also, it is tough to find High-quality datasets. Comparing the different outcomes is complicated because studies rely on different datasets and evaluation methods. Malware can detect when they are being examined in a sandbox environment and can change its behaviour to escape from being detected. This makes dynamic analysis less useful in such cases. [7]

Based on some evaluation that are done in past few years, modern methods of malware classification have now started combining ensemble and hybrid approaches which uses many techniques. These methods focus on boosting accuracy and handling complex infections more effectively. Deep learning and ensemble learning models shows best performance in spotting modern and advanced threats. But they also have some drawbacks like more complexity and cost. [8]

Overall, all the researches show that combining machine learning and dynamic analysis methods together is a promising way for classifying the malware. Many studies have also improved in detecting and classifying unknown malware. They have also reached a very high accuracy and performance. But some issues like the model generalization, processing cost and the dataset quality still appear. These issues show the need of performing more research in order to create such malware categorisation methods that are more effective, reliable, dependable and useful as compared to the existing ones.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

3. METHODOLOGY

This section gives a step-by-step method of the research methods that are performed. The main aim of this study is to use machine learning and dynamic analysis methods to categorise malware by examining its behaviour. The complete process is divided into different steps which are clearly explained below.

3.1. Data Collection

Collecting malware samples is the first step of our study. These samples have different types of malwares. This includes worms and ransomware and spyware and trojans. The assorted malware aims to teach the model different behaviours.
Some typical software samples which are safe are also collected with the malware. This helps the model understand the difference between harmful and safe programs. A balanced dataset is important to keep the model from becoming biased toward one group.

3.2. Dynamic Analysis Environment

Each malware sample operates in a sandbox and a controlled environment after gathering data. Even if the malware is harmful, it cannot damage the machine since this environment is separate from the real system.

Every sample's behaviour is closely observed throughout execution. The sandbox's tools capture several activities, including:

- System calls
- File creation and deletion
- Changes in system registry
- Network communication

This phase is crucial since it aids in comprehending the behaviour of malware under actual circumstances.

3.3. Feature Extraction

The next stage is to extract valuable features from the behaviour data after it has been gathered. In essence, features are crucial details that aid in the detection of malware.

For example, if a program is making too many unusual system calls or trying to connect to unknown networks, these can be considered as strong indicators of malware. These features are selected and converted into a structured format like numerical values so that they can be used by machine learning models.

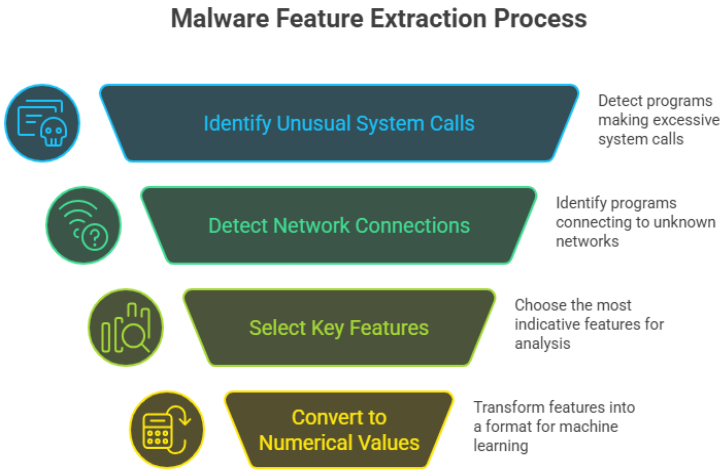

Fig. 3.3.1 Malware feature Extraction Process

3.4. Data Preprocessing

Before applying machine learning, the data needs to be cleaned and prepared. This step includes:

- Removing unnecessary or duplicate data
- Handling missing values

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

- Normalizing the data so that all features are on a similar scale

After preprocessing, the dataset is divided into two parts:

- Training data (used to train the model)
- Testing data (used to evaluate the model)

This helps in checking how well the model performs on new and unseen data.

3.5. Model Selection and Training

In this step, different machine learning algorithms are applied to the training data. Some commonly used algorithms in this research include:

- Decision Tree
- Random Forest
- Support Vector Machine (SVM)
- Naive Bayes

Every model looks for patterns in the data and attempts to determine the type of malware as well as whether a sample is malicious or not. Features and their appropriate labels are fed into the model during training.

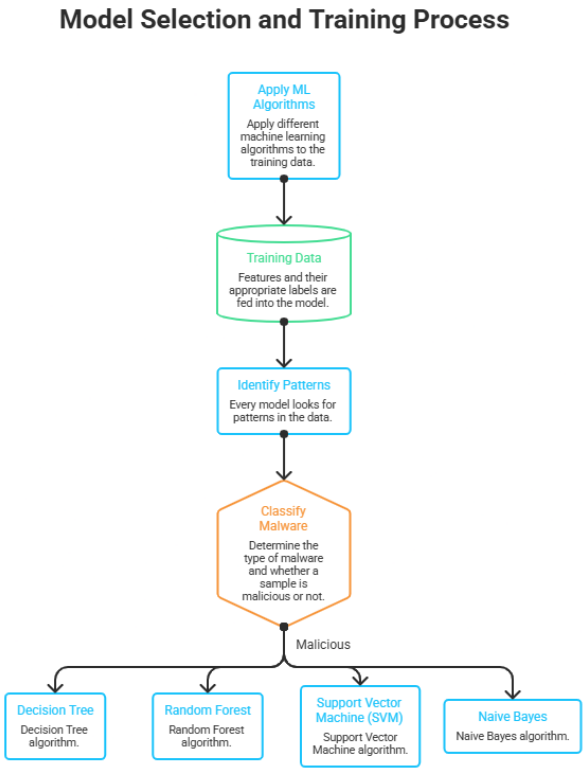

Fig.3.5.1 Model Selection and Training Process

3.6. Model Evaluation

After training, the models are tested using the testing dataset. This step helps in understanding how accurately the model can classify malware.

Different evaluation metrics are used, such as:

- Accuracy (overall correctness)
- Precision (correct positive predictions)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

- Recall (ability to detect malware)
- F1-score (balance between precision and recall)

These metrics give a clear idea of the performance of each model.

3.7. Comparison and Final Model Selection

Once all models are looked at and their results are compared the model which shows the best performance across all measures is chosen as the final one.

At the same time, if there are any problems found during the whole process like overfitting (a model which can only work on the training data) or long processing times are considered. Later work uses this approach to improve the model further more.

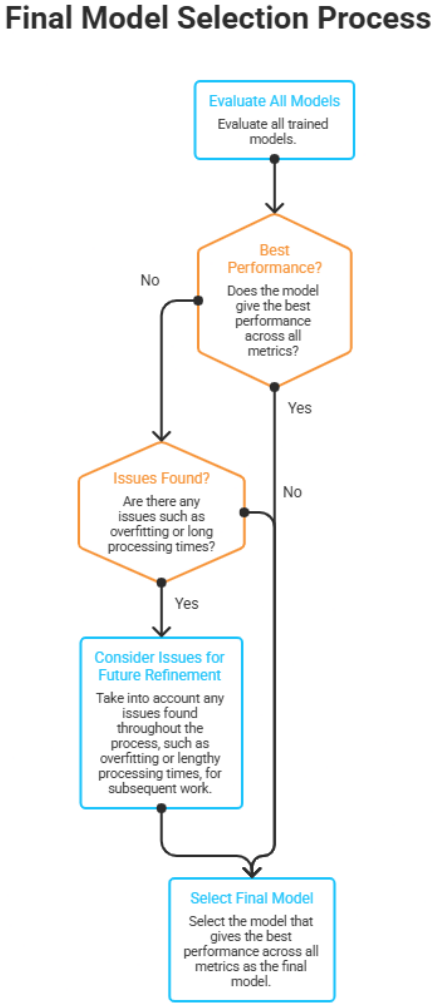

Fig.3.7.1 Final Model Selection Process

3.8. Summary of Method

This study begins with collecting malware data and then performing dynamic analysis on that data to monitor its behaviour and extracts attributes that are useful. After that, it uses machine learning to categorise the malware. Each step in the study is carefully planned to ensure that the model is accurate, reliable and it is able to detect new or modern malware.

This cautious approach helps to create an very effective malware categorisation system using the modern methods.

4. RESULTS

This part shows the outcome that we got by using machine learning and dynamic analysis methods for classifying malware. The result depends upon how the different models perform when they are evaluated on the prepared dataset.

4.1. Overview of Experimental Results

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

After training and testing the model, researchers have found that the machine learning models could effectively categorise malware based by monitoring its behaviour. Dynamic analysis boosts up the capability of the model by helping to detect malware activity in real time.
All of the models that were selected performed very well but there were noticeable differences present in their precision and reliability.

4.2. Performance of Machine Learning Models

Various different types of machine learning methods were used like Decision Tree, Random Forest, Support Vector Machine (SVM) and Naive Bayes. The results showed that:

- **Random Forest** method was the best among all the selected models as it managed the dataset very perfectly and also gave the highest accuracy. It also reduced the errors related to overfitting as compared to other methods.
- **Support Vector Machine (SVM)** also showed effectively a good result, especially for distinguishing benign and malicious data. But if we compare to earlier models then it is needed to be trained more.
- **Decision trees** were very easy to be understand and also delivered a good and efficient results but it was slightly less accurate as compared to Random Forests method. This method also included some signs of overfitting.
- **Naive Bayes** was the fastest model among all the models but it failed to match up the accuracy. Its performance for sample patterns was good but it lies behind for complex malware behaviour.

4.3. Evaluation Metrics

All the models were tested with standard performance measures which includes accuracy, precision, recall and F1-score.

- Accuracy: Random Forest produced the greatest results, with the majority of models achieving excellent accuracy.
- Precision: The models were able to accurately detect malware with little false alarms, demonstrating strong precision.
- Recall: The majority of the malware samples were successfully identified, according to high recall values.
- F1-score: Recall and precision were well-balanced, particularly for Random Forest and SVM.

The results showed that the models were good in identifying and classifying the malware.

4.4. Impact of Dynamic Analysis

One major finding of the study is that, how the dynamic analysis method improved the performance of classification by observing real-time behaviour of the malware, which includes system calls, network activities etc, the model's identified patterns are not visible in the static analysis.

This method helped in finding even those malware samples which try to hide their code using obfuscation techniques so that they are undetected.

4.5. Comparison with Traditional Methods

The proposed approach worked better as compared to the traditional signature-based detection methods. It performed well in detecting new and previously unknown samples and also known malware. This shows that combining dynamic analysis and machine learning works more effectively to tackle modern challenges of cybersecurity.

4.6. Limitations Observed

Even the results are promising but some limitations have also arrived during the experiment:

- Dynamic analysis required more time because each sample had to be executed in a sandbox.
- Some malware samples tried to hide their behaviour when running in a controlled environment.
- The effectiveness of model depended upon the quality and size of the data set.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

4.7. Summary

The outcomes showed that the method that was suggested has worked well in classifying the malware. Random Forest was the best method among all the other models and SVM was second among the others. The use of dynamic analysis improved the ability of the system to find complex and new or unknown malware.

These results supported the idea that the machine learning and dynamic analysis when combined together then it becomes a strong and very reliable method to identify and categorise the malware.

5. CONCLUSION

This research focused on classifying the malware by using machine learning and dynamic analysis methods. The main purpose was to understand that how the malware performs in a controlled environment(sandbox). Which further helps in training different machine learning models.

The result of this study clearly shows that the usefulness of the dynamic analysis in detecting malware by monitoring its actions in real time. Finding such patterns that shows harmful behaviour is easier when we start monitoring through the events like the system calls, file changes and network activities. This clearly shows that dynamic analysis method performs more efficiently than the older method that only rely on known signatures for detecting malwares.

The process of classification becomes more efficient with machine learning. Multiple models were tested among which Random Forest and Support Vector Machine excelled beyond other techniques. Both the models showed a strong ability in identifying unknown and new threats along with being highly accurate in classifying malware.

Using dynamic analysis and machine learning together provides a very strong and efficient method to address modern cybersecurity problems which is a important outcome of this study. This method reduces the risk of missing advanced malware from detection which uses modern hiding tricks and this method also improves the accuracy of malware detection.

Along with the advantage, this study also has some drawback. Creating a sandbox environment and running the malware sample in that controlled environment take a longer time. The amount and the quality of dataset directly affect the performance of machine learning models. These challenges shows that there is a room for improvement in the future research.

This study concludes by giving the outcome that the intelligent learning models when paired with behaviour-based analysis can enhance malware categorisation greatly. Researches in the future might focus on combining different methods to create detection system which can boost the productivity using larger datasets and are more reliable and faster

6. REFERENCES

[1] Berrios, S.; Leiva, D.; Olivares, B.; Allende-Cid, H.; Hermosilla, P. Systematic Review: Malware Detection and Classification in Cybersecurity. *Appl. Sci.* 2025, *15*, 7747. https://doi.org/10.3390/app15147747
[2] "Review of Contemporary Literature on Machine Learning based Malware Analysis and Detection Strategies", *GJCST*, vol. 16, no. E5, pp. 17–22, Mar. 2016, Accessed: Apr. 02, 2026. [Online]. Available: https://computerresearch.org/index.php/computer/article/view/1410
[3] Berrios, S.; Leiva, D.; Olivares, B.; Allende-Cid, H.; Hermosilla, P. Systematic Review: Malware Detection and Classification in Cybersecurity. *Appl. Sci.* 2025, *15*, 7747. https://doi.org/10.3390/app15147747
[4] https://www.atlantis-press.com/journals/ijcis/25899232/view
[5] Divyashree N and Nagaraja J, "Malware Classification with a Hybrid Deep Learning", *JoITML*, vol. 1, no. 1, pp. 18–21, Feb. 2024. https://qtanalytics.in/publications/index.php/JoITML/article/view/75
[6] arXiv:2307.14657 [cs.CR]
[7] Adel Abusitta, Miles Q. Li, Benjamin C.M. Fung, Malware classification and composition analysis: A survey of recent developments, Journal of Information Security and Applications,
Volume 59, 2021, 102828, ISSN 2214-2126, https://doi.org/10.1016/j.jisa.2021.102828
[8] https://www.researchgate.net/publication/393615508_Systematic_Review_Malware_Detection_and_Classification_in_Cybersecurity.