iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

# OPTIMIZED FEATURE SELECTION AND MACHINE LEARNING FOR ANDROID MALWARE IDENTIFICATION USING NEURAL NETWORKING WITH GENETIC ALGORITHM

**Nimayeela Abhinav, Rajamoni Pravalika, Sajjanapu Harikrishna,**
**Guide: Mrs. M Anusha**

Department of Electronics and Computer Engineering College, J.B. Institute of Engineering and Technology

**ABSTRACT**
The rapid growth of Android applications has led to an increasing number of malicious applications, posing serious threats to user privacy, data security, and system integrity. Traditional malware detection techniques, such as signature-based methods, are ineffective against newly emerging and obfuscated malware variants.
This paper presents an optimized Android malware detection system that integrates feature selection using Genetic Algorithms (GA) with Neural Network-based classification. The proposed system focuses on identifying the most relevant features from large-scale Android application datasets, thereby reducing dimensionality and improving detection accuracy.
The system employs a machine learning pipeline that includes data preprocessing, feature extraction, optimized feature selection using GA, and classification using Artificial Neural Networks (ANN). By selecting the most informative features, the model enhances performance while minimizing computational overhead.
Experimental results demonstrate that the proposed approach significantly improves detection accuracy, reduces false positives, and enhances model efficiency compared to traditional machine learning techniques. The integration of Genetic Algorithms and Neural Networks provides a robust and scalable solution for real-time Android malware detection.

## 1. INTRODUCTION

The Android operating system dominates the global smartphone market, making it a primary target for malware attacks. With millions of applications available across platforms, distinguishing between benign and malicious applications has become increasingly challenging.
Traditional malware detection techniques rely on static signatures, which fail to detect zero-day attacks and polymorphic malware. Moreover, the large volume of features extracted from Android applications often leads to high computational complexity and reduced model efficiency.
Machine learning-based approaches have emerged as a promising solution for malware detection. However, their effectiveness heavily depends on selecting relevant features from large datasets. Irrelevant or redundant features can negatively impact model accuracy and increase processing time.
To address these challenges, this paper proposes an optimized malware detection system that combines Genetic Algorithms for feature selection with Neural Networks for classification. The system aims to improve detection accuracy while reducing computational complexity by selecting only the most relevant features.
The proposed solution provides an efficient and scalable framework for Android malware detection, capable of adapting to evolving threats and supporting real-world deployment.

## 2. LITERATURE SURVEY

Android malware detection has been extensively studied using various machine learning and deep learning techniques. Early approaches focused on static analysis using permission-based features, API calls, and opcode sequences. While effective to some extent, these methods often suffer from high dimensionality and reduced scalability.

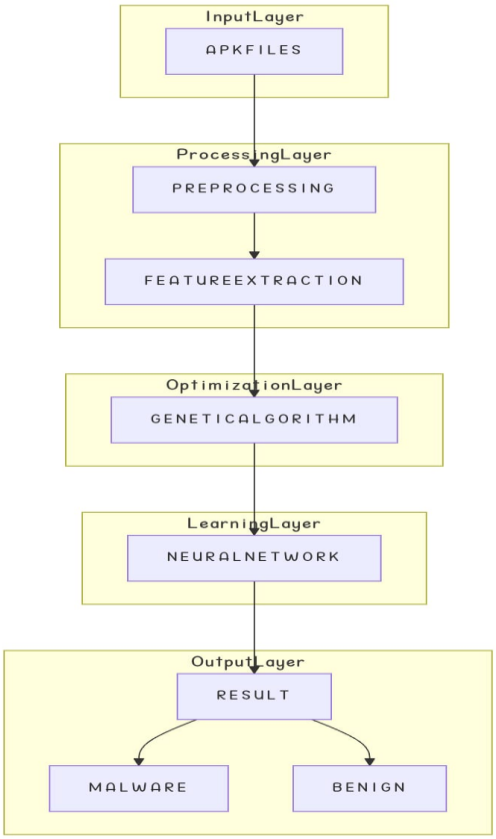
iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Dynamic analysis techniques were later introduced to monitor runtime behavior of applications. Although these methods improve detection accuracy, they are computationally expensive and require controlled environments.
Machine learning models such as Decision Trees, Support Vector Machines (SVM), and Random Forest have been widely used for malware classification. However, their performance is highly dependent on feature quality and selection.
Recent studies have explored deep learning approaches, particularly Artificial Neural Networks (ANN), for improved classification accuracy. However, these models often require large datasets and suffer from overfitting when irrelevant features are present.
Feature selection techniques such as Principal Component Analysis (PCA) and Information Gain have been used to reduce dimensionality. However, these methods may not always identify the optimal feature subset.
Genetic Algorithms (GA) have emerged as a powerful optimization technique for feature selection. By simulating natural evolution, GA efficiently identifies the most relevant features, improving both accuracy and efficiency.
Despite these advancements, there is still a lack of integrated systems that combine optimized feature selection with advanced neural network models for Android malware detection. The proposed system addresses this gap by integrating GA-based feature selection with ANN classification.

## 3. SYSTEM ARCHITECTURE

### A. Architecture Overview

The proposed system is designed as a machine learning-based architecture that processes Android application data through multiple stages, including preprocessing, feature selection, and classification.
The system takes input datasets (such as APK features or permission datasets) and processes them through a pipeline consisting of data cleaning, feature extraction, Genetic Algorithm-based optimization, and Neural Network classification.



iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

**B. System Components**

The system consists of the following key components:

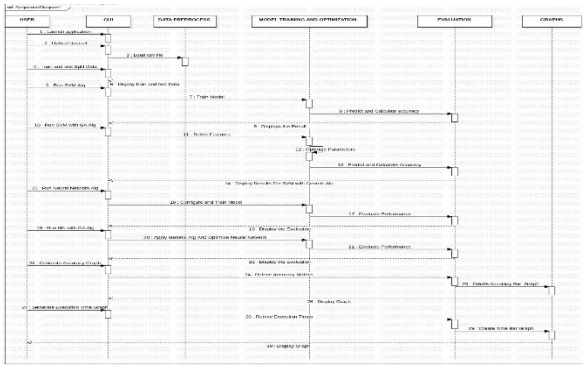
- **Data Collection Module:** Collects Android application datasets (benign and malicious samples).
- **Preprocessing Module:** Cleans and normalizes the data.
- **Feature Selection Module (GA):** Applies Genetic Algorithm to select optimal features.
- **Classification Module (ANN):** Uses Neural Networks to classify applications.
- **Evaluation Module:** Measures accuracy, precision, recall, and F1-score.

The fitness function in the Genetic Algorithm evaluates feature subsets based on classification accuracy.

**C. System Workflow**

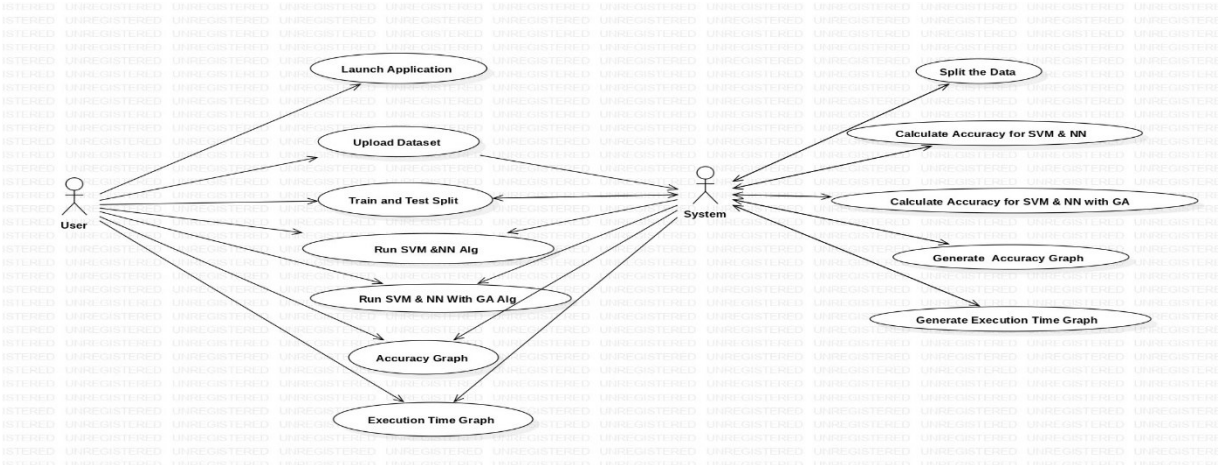
The system follows a structured workflow:

1. Dataset collection
2. Data preprocessing
3. Feature extraction
4. Genetic Algorithm-based feature selection
5. Neural Network training
6. Malware classification



**D. Use Case and Detection Flow**

The system allows users or analysts to input application data and receive classification results indicating whether the application is benign or malicious. The detection flow includes feature optimization followed by classification.



iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

## 4.Implementation

### A. Data Collection and Preprocessing

Datasets such as Drebin or other Android malware datasets are used. Data preprocessing includes handling missing values, normalization, and feature encoding.

### B. Genetic Algorithm for Feature Selection

The Genetic Algorithm is used to select the most relevant features. It involves:

- Population initialization
- Fitness evaluation
- Selection
- Crossover
- Mutation

This process iteratively improves feature subsets to maximize classification accuracy.

### C. Neural Network Classifier

An Artificial Neural Network (ANN) is used for classification. It consists of:

- Input layer (selected features)
- Hidden layers (processing)
- Output layer (malware/benign classification)

The model is trained using labeled data and optimized using backpropagation.

### D. API and Data Flow

The system processes input data through the pipeline and outputs classification results. Each stage communicates seamlessly to ensure efficient data flow.

### E. Execution Workflow

1. Load dataset
2. Preprocess data
3. Apply GA for feature selection
4. Train Neural Network
5. Evaluate model
6. Predict malware

## 5. Testing and Validation

### A. System Performance

The system is tested for accuracy and efficiency using multiple datasets. Results show improved performance due to optimized feature selection.

### B. Feature Optimization Accuracy

The Genetic Algorithm successfully reduces feature space while maintaining high classification accuracy.

### C. Model Stability

The Neural Network provides consistent results across different datasets and conditions.

### D. Overall Performance

The combined GA + ANN model achieves high accuracy, reduced false positives, and efficient processing.

**Table I: Evaluation Summary**

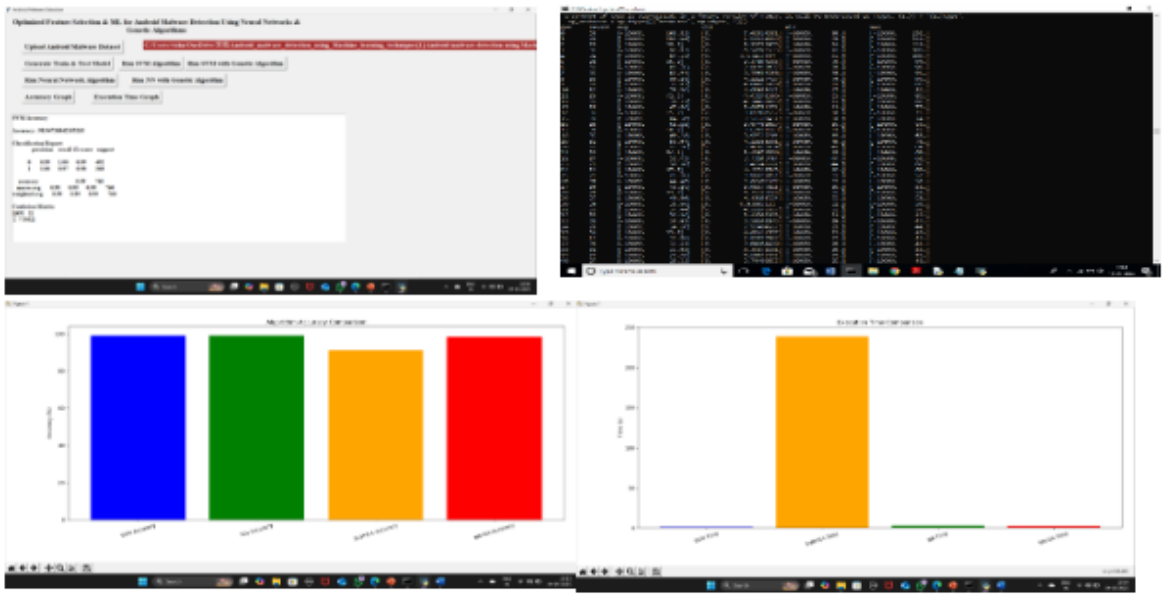
| Metrics | Observation |
|---|---|
| Accuracy | High |
| Feature Reduction | Significant |
| False Positives | Reduced |

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

| Metrics | Observation |
|---|---|
| Processing Time | Optimized |
| Model Stability | Consistent |

**6. Experimental Results**

- Visualization of dataset distribution
- Feature selection comparison
- Accuracy graphs
- Confusion matrix
- Training vs testing performance



## 7. CONCLUSION

The proposed system demonstrates an effective approach for Android malware detection by combining Genetic Algorithm-based feature selection with Neural Network classification.

By optimizing feature selection, the system reduces computational complexity while improving detection accuracy. The integration of machine learning techniques enables efficient identification of malicious applications, even in the presence of evolving threats.

The results confirm that the proposed method outperforms traditional approaches and provides a scalable solution for real-world deployment.

**Future Enhancements**

- Integration with deep learning models (CNN, LSTM)
- Real-time malware detection system
- Cloud-based deployment
- Hybrid static + dynamic analysis
- Explainable AI for better interpretability

## 8. REFERENCES

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

1) N. Xie, Z. Qin, and X. Di, "GA-StackingMD: Android Malware Detection Method Based on Genetic Algorithm Optimized Stacking," *Applied Sciences*, vol. 13, no. 4, p. 2629, 2023. [Online]. Available: https://www.mdpi.com/2076-3417/13/4/2629
2) M. Ş. Beştaş and Ö. B. Dinler, "Detection of Android Based Applications with Traditional Metaheuristic Algorithms," *International Journal of Pure and Applied Sciences*, vol. 9, no. 4, pp. 381–392, 2023. [Online]. Available: https://dergipark.org.tr/en/pub/ijpas/article/1382344
3) A. Aljanabi, A. A. Mosa, and A. A. A. Jassim, "Android Malware Detection Using Machine Learning with Feature Selection Based on Genetic Algorithm," *Mathematics*, vol. 9, no. 21, p. 2813, 2021. [Online]. Available: https://www.mdpi.com/2227-7390/9/21/2813
4) M. A. Pathan, M. A. N. Patil, and S. T. Bagwan, "Android Malware Detection Using Genetic Algorithm-Based Feature Selection," in *Proceedings of the International Conference on Emerging Trends in Engineering*, Springer, 2022, pp. 231–239. [Online]. Available: https://link.springer.com/chapter/10.1007/978-981-99-7954-7_19
5) V. Thakur and M. Sharma, "Android Malware Detection Using Genetic Algorithm-Based Optimized Feature Selection and Deep Learning," *International Journal of Innovative Research in Technology*, vol. 9, no. 11, pp. 456–460, 2023. [Online]. Available: https://ijirt.org/publishedpaper/IJIRT159655_PAPER.pdf
6) R. Mounika, M. Sireesha, and N. Shireesha, "Android Malware Detection Using Genetic Algorithm-Based Optimized Feature Selection and Machine Learning," *Journal of Emerging Science and Technology*, vol. 15, no. 2, pp. 145–151, 2022. [Online]. Available: https://jespublication.com/uploads/2024-V15I02020.pdf