Image-Based Malware Classification Using Deep Learning Models

Dipta Dhor, Sujoy Sarkar, Sadhon Kumar Dev, Saikat Sarkar
Department of Computer Science and Engineering
International Islamic University Chittagong (IIUC), Bangladesh
Emails: diptadhor2002@gmail.com, sujoy7291@gmail.com, skdsadhon@gmail.com, sarkarsaikat8526@gmail.com

***Abstract*—Malware detection has become increasingly challenging due to the rapid evolution of malicious software and the limitations of traditional signature-based detection techniques. This paper presents an image-based malware classification framework using deep learning models. Executable binaries are transformed into grayscale images, enabling malware classification as a computer vision problem. A dataset consisting of 29 classes, including benign software and multiple malware families, is used to evaluate three models: a custom Convolutional Neural Network (CNN), MobileNetV2, and Vision Transformer (ViT). To address severe class imbalance, class-weighted loss functions are applied during training. While MobileNetV2 utilizes pretrained weights through transfer learning, the Vision Transformer is trained from scratch due to the unavailability of pretrained weights. Experimental results demonstrate that MobileNetV2 achieves superior performance with a macro F1-score of 91.33%, highlighting the effectiveness of transfer learning for robust malware classification.**

***Index Terms*—Malware Classification, Image-Based Analysis, Deep Learning, MobileNetV2, Cybersecurity**

I. Introduction

Malware poses a serious threat to modern computing systems due to its increasing volume, diversity, and sophistication. Traditional detection techniques such as signature-based methods are ineffective against zero-day and polymorphic malware. As a result, intelligent malware detection mechanisms based on machine learning and deep learning have gained significant attention.

Image-based malware classification is an effective approach in which binary executables are visualized as grayscale images. This representation preserves the structural characteristics of binaries and enables the use of deep learning models originally designed for computer vision tasks. This paper investigates and compares convolutional and transformer-based models for multi-class malware image classification.

II. Related Work

Nataraj et al. introduced malware visualization by converting binary files into grayscale images and demonstrated effective classification using image processing techniques. Subsequent research applied convolutional neural networks to automatically extract discriminative features from malware images.

Recent studies have explored transfer learning using pretrained CNN architectures to improve classification accuracy. However, limited work has investigated the effectiveness of Vision Transformer models for malware image classification. This work provides a comparative analysis of CNN-based and transformer-based models under highly imbalanced dataset conditions.

III. Methodology

A. Dataset Description

The dataset used in this study consists of malware and benign executable files converted into grayscale images. Benign samples include legitimate DLL, EXE, MUI, and other verified system files. Malware samples span multiple families, resulting in a 29-class classification problem.

TABLE I
Dataset Summary

Description	Value
Total images	19,209
Benign images	9,870
Malware images	9,339
Number of classes	29
Training samples	15,379
Validation samples	3,830
Image size	224×224

B. Preprocessing

Binary files were read as byte sequences and converted into grayscale images by mapping byte values (0–255) to pixel intensities. All images were normalized to the range [0,1]. The dataset was split into training and validation sets using an 80:20 ratio.

C. Class Imbalance Handling

The dataset exhibits severe class imbalance. To address this issue, class weights were computed using a balanced weighting strategy and applied during model training to prevent bias toward majority classes.

D. Model Architectures

Three deep learning models were evaluated:

Custom CNN: A baseline convolutional architecture designed to learn spatial features from malware images.

MobileNetV2: A lightweight pretrained CNN using depthwise separable convolutions and transfer learning with ImageNet weights.

Vision Transformer (ViT): A transformer-based model that processes images as fixed-size patches using self-attention mechanisms. In this study, the Vision Transformer was trained from scratch without pretrained weights due to download and compatibility limitations, which impacted its generalization performance on the given dataset.

IV. Evaluation Metrics

Model performance was evaluated using accuracy, macro-averaged precision, recall, and F1-score. Macro averaging was selected to ensure equal importance for all classes.

A. Accuracy

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

B. Precision

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

C. Recall

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

D. F1-score

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

V. Results

A. Quantitative Evaluation

Table II presents the macro-averaged performance metrics of all models.

TABLE II
Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-score
Custom CNN	0.7279	0.5897	0.8024	0.6387
MobileNetV2	**0.9708**	**0.9013**	**0.9283**	**0.9133**
Vision Transformer	0.5235	0.5032	0.7311	0.5447

B. Confusion Matrix

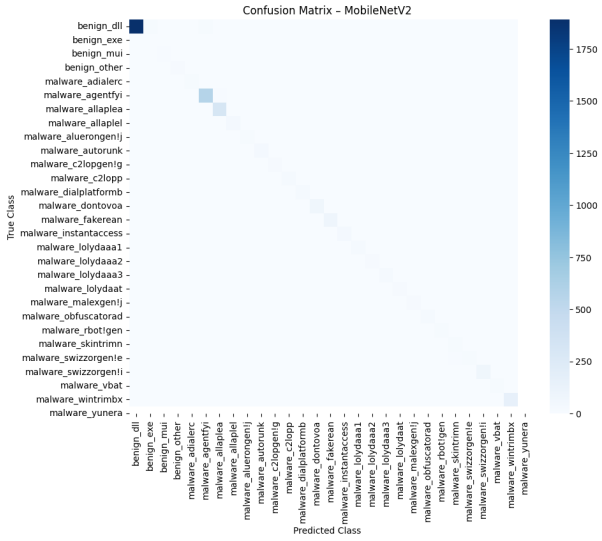

Fig. 1. Confusion Matrix of MobileNetV2

C. ROC–AUC Analysis

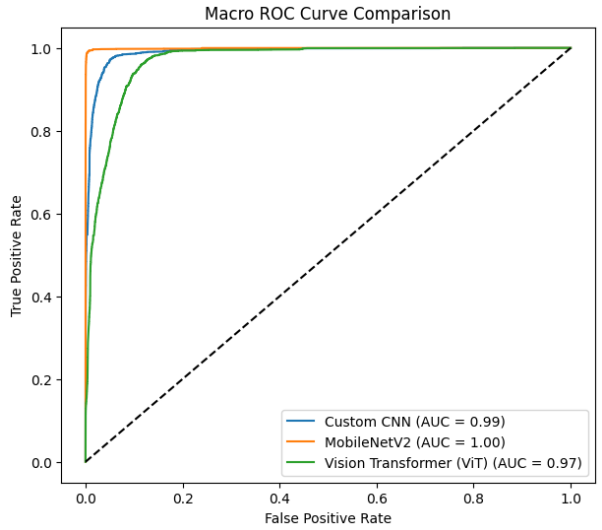

Fig. 2. Macro ROC–AUC Curve Comparison

D. Performance Comparison

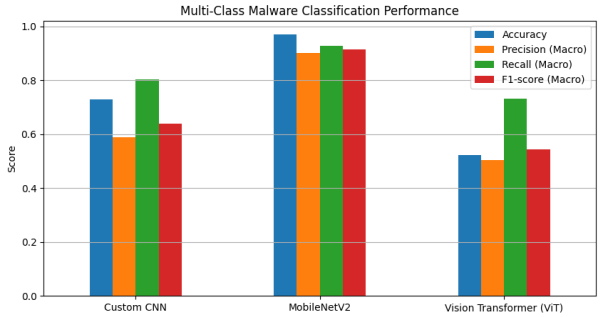

Fig. 3. Performance Comparison Bar Chart

VI. Discussion

The experimental results clearly demonstrate that MobileNetV2 outperforms the custom CNN and Vision Transformer models across all evaluation metrics. The use of transfer learning enables MobileNetV2 to extract robust and discriminative features despite severe class imbalance. In contrast, the Vision Transformer was trained from scratch without pretrained weights, which significantly limited its ability to learn effective representations from the available dataset, leading to lower overall performance.

VII. Conclusion

This paper presented an image-based malware classification framework using deep learning models. Experimental evaluation confirms that MobileNetV2 provides superior performance for multi-class malware classification. Future work will explore pretrained transformer models and hybrid CNN-transformer architectures to improve generalization.

References

[1] L. Nataraj et al., "Malware Images: Visualization and Automatic Classification," IEEE VizSec, 2011.
[2] J. Kiger et al., "Malware Binary Image Classification Using CNNs," ICCWS, 2022.
[3] Kaggle, "Malware Classification Dataset (MalImg)."