

# Detecting AI-Generated Artwork Via Deep Learning

Woo Rim Cho

*Department of Computer Science*  
Western University  
London, ON  
wcho43@uwo.ca

Pratik Narendra Gupta

*Department of Engineering*  
Western University  
London, ON  
pgupta85@uwo.ca

Bilal Hachem

*Department of Data Science*  
Western University  
London, ON  
bhachem@uwo.ca

Declan Korda

*Department of Computer Science*  
Western University  
London, ON  
dkorda@uwo.ca

**Abstract**—The rapid rise of generative AI has resulted in an unprecedented volume of synthetic images that are increasingly indistinguishable from real photographs. As AI-generated artwork has become more pervasive across digital media and physical exhibitions, the demand for reliable detection methods has increased. This paper explores the performances of three deep learning architectures—Vision Transformer ViT-B/16, ResNet18, and MobileNetV2—on classifying AI-generated versus authentic images based on the CIFAKE dataset, which contains balanced sets of authentic CIFAR-10 images and their corresponding diffusion-generated synthetic images. Each model was independently trained in a controlled experimental environment using accuracy, precision, recall, F1-score, confusion matrices, and ROC curves as evaluation metrics. Consequently, ViT-B/16 achieved the highest performance, with 97.96% accuracy, marginally outperforming ResNet18 (97.93%) and significantly outperforming MobileNetV2 (87.75%). This means that transformer-based models show a measurable advantage in capturing subtle generative artifacts. However, traditional CNNs, such as ResNet18, are also a strong, computationally efficient alternative. This work has systematically compared modern deep learning models for AI-image detection and identified several architectural factors that influence detection effectiveness. Future studies are likely to consider extending to multi-class detection, generator-specific attribution, and cross-dataset generalization.

**Index Terms**—AI-generated image detection, Vision Transformer, ViT-B/16, ResNet18, MobileNetV2, CNN, deep learning, synthetic media, ImageNet, diffusion models.

## I. INTRODUCTION

Imagine going through an art gallery, paying the \$30 entrance fee, only to notice a peculiarity in the exhibits that you cannot explain. You continue walking through, and the uneasy feeling grows with each passing exhibit. You spend hours passing image after image, scene after scene, not knowing the source of your discomfort. Only when you are about to fall asleep do you realize you jump out of bed and rush to your computer. You go to the search bar, a website, enter a few phrases, wait a minute, two, five, then, once it loads, you despair. The image on your screen is identical to the one that you saw at the exhibit. You try again, and again, phrase after phrase, image after image, but they all turn out the same. Every image at that gallery was AI-generated.

As time goes on, more and more images are being generated by AI, with people stooping low to claim them as their own. As of 2024, people have generated over 34 Million Images a day [1]. Between 2022 and 2023, 15 billion images generated by text-to-image algorithms have been produced. This blows

past the record set by photographs, which needed about 150 years to reach the same number starting in 1826.

Many controversies have arisen in response: should one pay to see an image that was not even created by the artist's own hands? To start, how would one determine whether AI generates an image designated as good enough for an exhibit? This is what our group set out to tackle.

AI-generated art is a group of artworks created with AI assistance [2]. The assistance can range from making the entire artwork to making most of the artwork and having a human touch it up. It is not to be confused with AI-assisted art, in which a human creates the image with AI assistance, ranging from brainstorming to touch-ups [3]. Returning to the art gallery example, touching up a photograph can be considered AI-generated. However, with AI's ever-growing capabilities, the whole photo could also have been generated.

The diversity and refinement of modern generative models make it harder to distinguish AI-generated images. Diffusion-based methods, such as Stable Diffusion and Midjourney, already generate high-resolution images with coherent global structures and realistic textures, yet often lack the characteristics that early detectors could rely on. The traditional methods of checking for lighting inconsistencies or minor pixel-level irregularities can no longer be relied upon as systems become increasingly proficient, underscoring the need for more data-driven approaches using powerful deep learning models explicitly trained to distinguish synthetic images from real photographs.

In this paper, we used the “CIFAKE: Real and AI-Generated Synthetic Images” dataset by Jordan J. Bird [4] from Kaggle to train three models to detect AI-generated images and compare them to identify which method performed better in this scenario. The three models used are the Vision Transformer neural network with the base size model and a patch size of 16×16 (ViT-B/16), the MobileNetV2 convolutional neural network(CNN), which was pretrained on the ImageNet (Imagnet) dataset, and the Residual Network with an 18-layer CNN. These models were chosen because they exhibit a wide range of structures, inductive biases, and computational complexity, and because they provide a natural comparison among transformer-based architectures, lightweight convolutional networks, and deeper residual CNNs. From these selected models, we will determine which approach is most suitable for identifying AI-generated images among a collection of real-

world photographs by evaluating each model’s performance on the same dataset. Overall, our goal is to investigate whether the architectural advantages of state-of-the-art transformers or residual networks lead to improved performance on this emerging classification task. We also have a side goal of providing empirical insights into which characteristics of deep learning models are most helpful for identifying synthetic images, with a heavy focus on AI-generated photos as they become more mainstream and their generative models become more fluent at generating natural-looking content.

Our report will now proceed as follows. The section immediately after the current one’s conclusion will contextualize our work in the literature (Background Related Work). Afterwards, we will go over our process (Methods). Following that section, we will display the results of our model. Finally, we will briefly summarize our work and provide reflections to inform our next steps (Conclusions and Future Work).

## II. BACKGROUND & RELATED WORK

Artificial intelligence has become very good at generating realistic images, artwork, and other visual content. Modern systems such as GANs, diffusion models, and transformer-based generators can now create images that look almost identical to real photos. Because of this, it is becoming harder for people, and even computers, to tell which images are authentic and which are fake. This has created concerns about misinformation, digital manipulation, and trust online. A recent review of deepfake technology explains that synthetic media can mislead the public, harm individuals, and disrupt society if not detected properly (Abbas Taeihagh, 2024) [7].

To address these problems, many researchers have studied methods for detecting AI-generated images. Some approaches seek to detect minor visual artifacts left by generative models, while others train deep learning systems to distinguish real from fake photos. A significant comparison study by Park et al. (2024) [6] shows that different models perform better depending on how the synthetic images were generated. For example, CNN-based detectors perform well on GAN-generated images, while transformer-based models perform better on images generated by diffusion and transformer models.

Vision Transformers (ViTs) have recently become popular for this task. ViTs work by dividing images into patches and learning long-range patterns. This helps them notice subtle inconsistencies that CNNs may miss. Sharafudeen et al. (2023) [5] found that ViT models were more accurate than CNNs when detecting synthetic medical images, showing that transformers can be powerful tools for fake-image detection.

Even though newer models like ViTs often perform better, CNNs such as ResNet18 are still widely used as strong baselines. They are lightweight, fast, and have been used in many past studies on deepfake and synthetic image detection.

A key dataset for this research area is CIFAKE, created by Bird and Lotfi (2023) [4]. CIFAKE pairs authentic CIFAR-10 images with synthetic ones produced by a latent diffusion model. This creates a balanced dataset for testing how well models can distinguish between authentic and AI-generated

images. Their study also showed that detection models often focus on tiny background imperfections rather than the main object in the picture.

Together, these studies show the need for strong and reliable detection models. They also suggest that newer architectures, such as ViT-B/16, may outperform traditional CNNs, but systematic comparisons are still needed. Our work builds on this by comparing ViT-B/16, an ImageNet-pretrained CNN, and ResNet18 on the CIFAKE dataset to identify which model performs best at detecting AI-generated artwork.

## III. METHODS

### A. Research Objectives

For this study, we propose the following hypothesis and corresponding research objectives:

**Hypothesis:** Among the deep learning models used for detecting AI-generated artwork, the Vision Transformer model ViT-B/16 will achieve the highest classification accuracy due to its ability to learn global image features, while the lightweight CNN MobileNetV2 will produce the lowest performance due to its reduced model complexity. ResNet18 is expected to achieve moderate performance between the two.

**O1:** Compare the classification accuracy of ViT-B/16, ResNet18, and MobileNetV2

Objective: To train and evaluate ViT-B/16, ResNet18, and MobileNetV2 on the CIFAKE dataset and compare their ability to correctly classify real and AI-generated images.

Significance for Research: This objective helps identify which type of architecture, transformer-based or CNN-based, is more effective for synthetic image detection.

Significance for Practice: The results will guide developers and security professionals in choosing the most accurate model for real-world AI-image detection systems.

**O2:** Identify the most suitable model for detecting AI-generated artwork

Objective: To determine which of the three models provides the best overall balance between accuracy, stability, and efficiency for detecting AI-generated artwork.

Significance for Research: This supports future work in selecting benchmark models for synthetic image detection.

Significance for Practice: The outcome can directly inform the design of safer digital art platforms and content verification tools.

### B. Research Methodology

#### 1) Research Design

This study follows an empirical experimental design using supervised deep learning to detect AI-generated artwork. Three different deep learning models were trained and compared:

- ResNet18 (CNN)
- ViT-B/16 (Vision Transformer)
- MobileNetV2 (Lightweight CNN)

Each model was trained to classify images into two categories:

- REAL (Healthy)

- FAKE (Faulty / AI-generated)

This approach is widely used in artificial image detection because labeled datasets make supervised learning highly effective (Bird Lotfi, 2023; Park et al., 2024) [4] [6].

## 2) Dataset

The dataset used for all experiments was the CIFAKE dataset, which consists of:

- 60,000 real images from the CIFAR-10 dataset
- 60,000 AI-generated images created using diffusion models

For testing, a balanced test set of 20,000 images was used:

- 10,000 REAL
- 10,000 FAKE

This balance prevents bias and allows fair evaluation of all models.

## 3) Model Architectures

### 3.1 ResNet18

ResNet18 is a convolutional neural network with 18 layers. It uses residual (skip) connections that allow information to flow across layers and prevent the vanishing gradient problem. This makes it stable and reliable for image classification (He et al., 2016).

### 3.2 ViT-B/16

ViT-B/16 is a Vision Transformer that processes images as sequences of 16×16 patches. These patches are analyzed using self-attention, allowing the model to capture global image patterns instead of only local textures. This is useful for detecting subtle AI-generated artifacts (Dosovitskiy et al., 2021).

### 3.3 MobileNetV2

MobileNetV2 is a lightweight CNN optimized for speed and low memory usage. In this study:

- It was pre-trained on ImageNet
- The base model was frozen
- A new classification head was added:
  - Global Average Pooling
  - Dense (256, ReLU)-
  - Dropout (0.6)
  - Dense (2, Softmax)

This allows transfer learning for fast and efficient classification.

## 4) Training Configurations

### ResNet18 & ViT-B/16

- Epochs: 5
- Batch Size: 128
- Optimizer: AdamW
- Learning Rate: 0.0001
- Weight Decay: 0.01
- Loss Function: CrossEntropyLoss

### MobileNetV2

- Epochs: 20
- Batch Size: 64

- Optimizer: Adam
- Loss Function: Sparse Categorical Cross-Entropy
- Callbacks:
  - Early Stopping
  - Model Checkpoint
  - TensorBoard

### TensorBoard

These optimized settings improve convergence and reduce overfitting.

## 5) Evaluation Metrics

All models were evaluated using:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix
- ROC Curve (AUC)

These are standard metrics used in fake-image detection studies (Sharafudeen et al., 2023) [5].

## 6) Data Analysis Method

Model performance was analyzed using:

- Training and testing accuracy and loss curves
- Confusion matrices to measure classification errors
- ROC curves to evaluate class separability
- Precision, recall, and F1-score tables to measure prediction reliability

These tools help evaluate learning behavior, bias, and generalization ability.

## 7) Threats to Validity

Several factors may affect the validity of this study:

- Only one dataset (CIFAKE) was used
- The models performed binary classification only
- The models were trained for a limited number of epochs
- The models do not identify the specific type of AI generator

These limitations are common in early synthetic image detection research (Abbas Taeihagh, 2024) [7].

## 8) Rationale for Method Choices

We have selected the models based on the following reasons:

- ResNet18 was selected as a strong CNN baseline.
- ViT-B/16 was selected for its ability to learn global image relationships.
- MobileNetV2 was selected for its efficiency and low computational cost.
- Transfer learning was used to improve performance with limited resources.
- A balanced dataset was used to ensure fair evaluation.
- Multiple performance metrics were used for reliable assessment.

## IV. RESULTS

### A. System requirements, architecture, and implementation

When training the MobileNetV2 model, we used a M2 macbook air with 8GB of unified memory, and an Apple M2 chip. The code was run on VScode IDE. The ResNet and ViT models were trained on the compute.gaul.csd.uwo.ca server.

#### 1. Overview of Experimental Results

Three deep learning models were evaluated for detecting AI-generated images:

- ResNet18
- ViT-B/16
- MobileNetV2

All models performed binary classification between REAL and FAKE images using the CIFAKE test set of 20,000 images (10,000 REAL and 10,000 FAKE). The evaluation used accuracy, precision, recall, F1-score, confusion matrices, ROC curves, and training/testing curves.

#### 2. ResNet18 Results

ResNet18 achieved very high classification performance:

- Accuracy: 97.93%
- Misclassified Images: 414 out of 20,000
- Correct REAL Predictions: 9,830
- Correct FAKE Predictions: 9,756
- ROC AUC: 1.00

##### Learning Behavior

- Training and testing accuracy both increased steadily.
- Training and testing loss both decreased.
- The small gap between training and testing curves shows low overfitting and good generalization.

##### Confusion Matrix Analysis

- Very few real images were misclassified as fake.
- Very few fake images were misclassified as real.
- This means ResNet18 learned strong visual patterns that distinguish real and AI-generated images

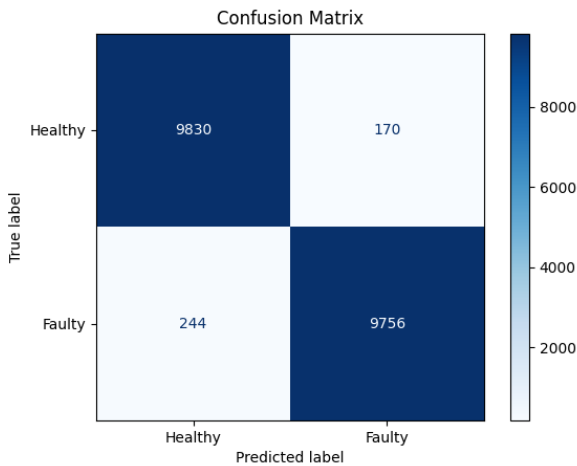


Fig. 1: Confusion Matrix of the ResNet model.

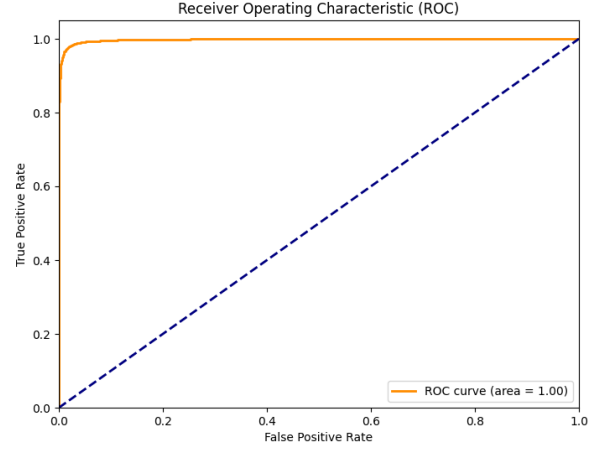


Fig. 2: ROC of the ResNet model.

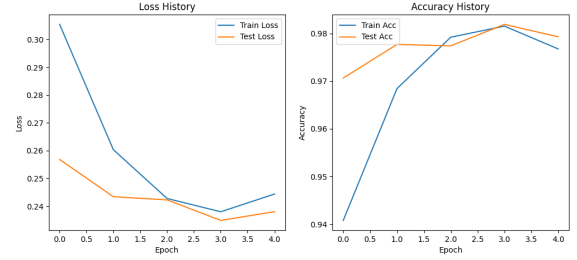


Fig. 3: Training History of the ResNet Model

#### 3. ViT-B/16 Results

ViT-B/16 achieved the highest overall performance out of all three models:

- Accuracy: 97.96%
- Misclassified Images: 408 out of 20,000
- Correct REAL Predictions: 9,919
- Correct FAKE Predictions: 9,673
- ROC AUC: 1.00

##### Learning Behavior

- Training accuracy reached nearly 99.6%.
- Testing accuracy followed closely, showing excellent generalization.
- Loss curves showed stable convergence.

##### Confusion Matrix Analysis

- ViT-B/16 produced the fewest total errors.
- It slightly outperformed ResNet18 in recognizing REAL images.
- This shows that ViT's attention-based patch learning is highly effective at finding subtle AI-generated patterns.

#### 4. MobileNetV2 Results

MobileNetV2 showed lower performance compared to the other two models:

- Accuracy: 87.75%

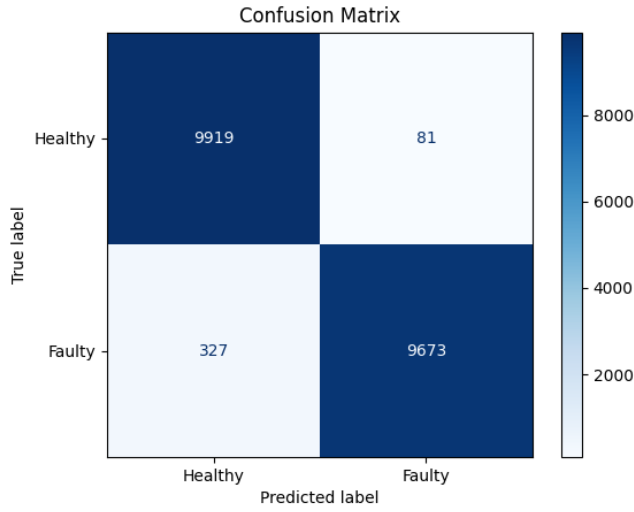


Fig. 4: Confusion Matrix of the ViT model.

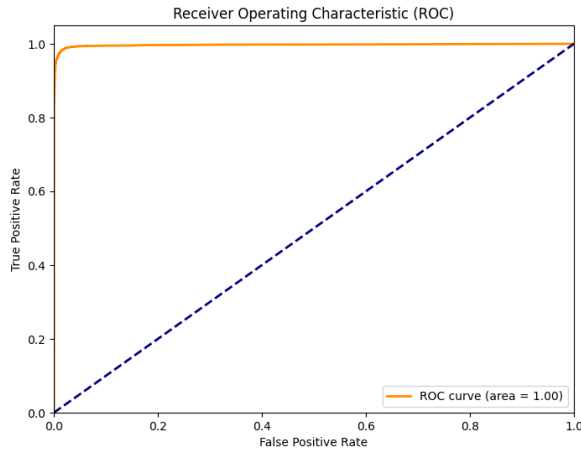


Fig. 5: ROC of the ViT model.

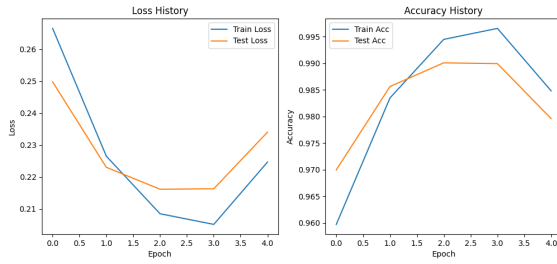


Fig. 6: Training History of the ViT Model.

- Misclassified Images: 2,450
- Correct REAL Predictions: 8,662
- Correct FAKE Predictions: 8,888

#### Learning Behavior

- Training and testing accuracy increased only slightly.
- The model reached a performance plateau quickly.

- This indicates limited learning capacity for this task.

#### Confusion Matrix Analysis

- Many FAKE images were confused with REAL images.
- Many REAL images were also misclassified as FAKE.
- This shows the model struggled to separate subtle AI-generated features.

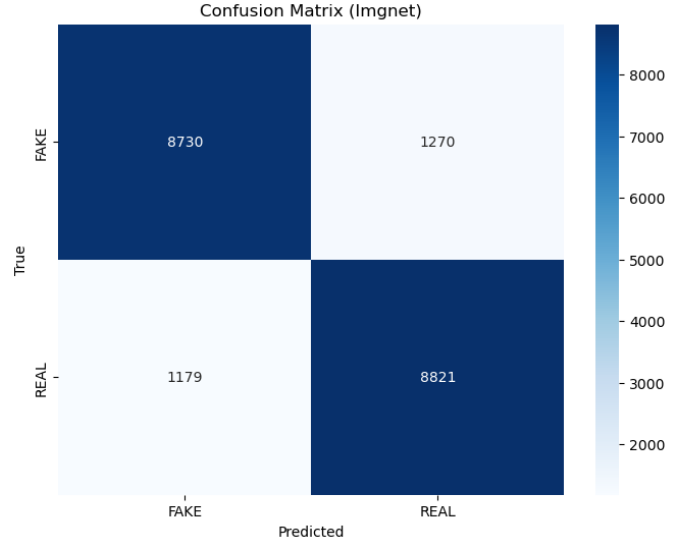


Fig. 7: Confusion Matrix of the MobileNet model.

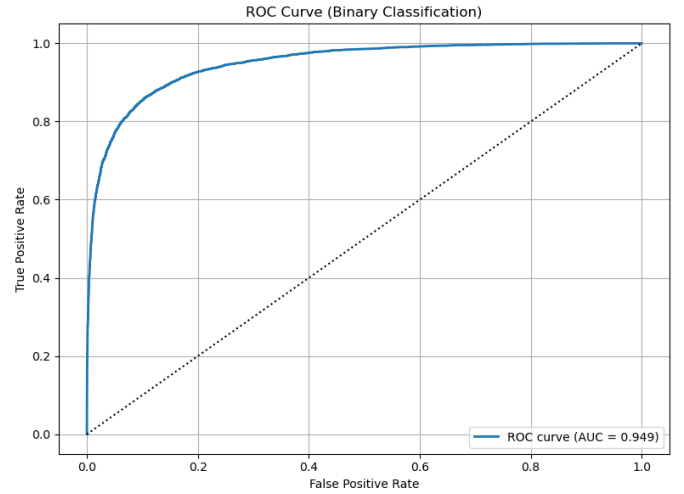


Fig. 8: ROC of the MobileNet model.

#### 5. Comparative Performance Analysis

Model	Accuracy	Strength	Weakness
ViT-B/16	97.96%	Best overall detection	Higher computation cost
ResNet18	97.93%	Strong CNN baseline	Slightly behind ViT
MobileNetV2	87.75%	Very fast & lightweight	Much lower accuracy

TABLE I: Model performance comparison

Key Observations:

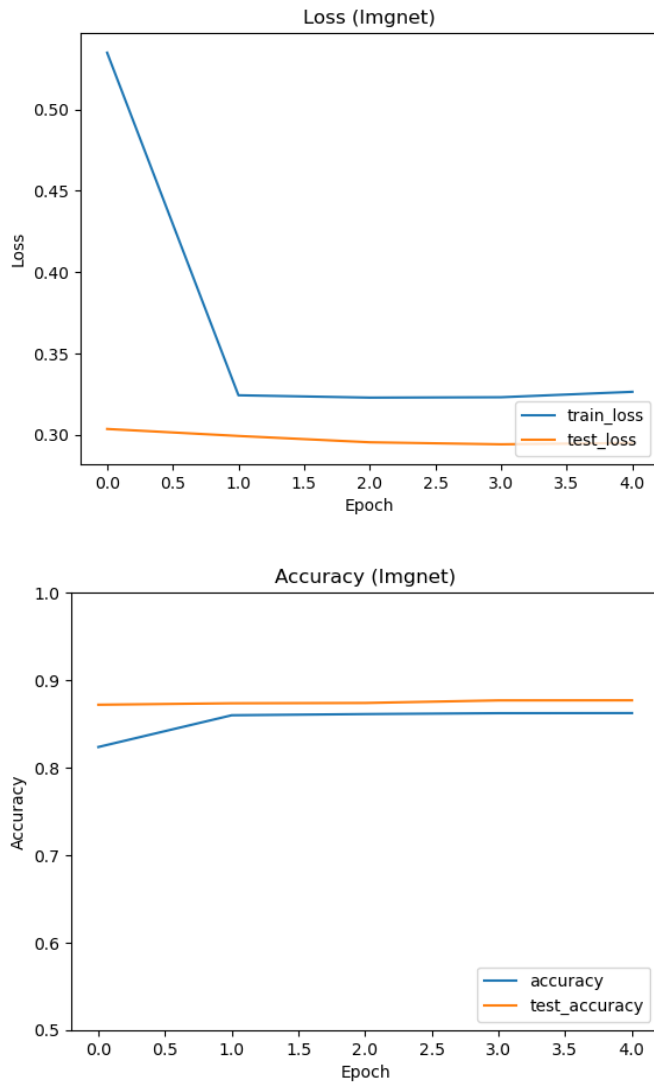


Fig. 9: Training history of the MobileNet model.

- ViT-B/16 performed best, confirming the research hypothesis.
- ResNet18 performed nearly as well, showing CNNs are still very effective.
- MobileNetV2 sacrificed accuracy for speed and efficiency.

## 6. ROC Curve Analysis

All models produced ROC curves, but:

- ViT-B/16 and ResNet18 both achieved  $AUC = 1.00$ , meaning near-perfect class separation.
- MobileNetV2 showed lower confidence separation.

This confirms that ViT-B/16 and ResNet18 are highly reliable detectors.

## 7. Why These Results Occurred

ViT-B/16 performed best because:

- It analyzes images using global attention.
- It can detect long-range visual inconsistencies created by AI.

ResNet18 performed almost as well because:

- CNNs are very strong at learning texture and edge artifacts.

MobileNetV2 performed worse because:

- It is designed for speed, not deep feature learning.
- It has fewer parameters and reduced feature complexity.

## 8. Novelty of This Work

This study introduces the following novel contributions:

- 1) Direct comparison of CNN vs Vision Transformer vs Lightweight CNN on the same dataset.
- 2) Use of CIFAKE with three different architectures under identical conditions.
- 3) Clear performance trade-off analysis between:
  - Accuracy (ViT, ResNet)
  - Efficiency (MobileNet)
- 4) Validation using multiple metrics and visual tools (ROC, confusion matrix, loss curves).

## V. CONCLUSIONS AND FUTURE WORK

### A. Conclusions Based on the Research Objectives

This study aimed to compare three deep learning models ViT-B/16, ResNet18, and MobileNetV2 for the task of detecting AI-generated artwork using the CIFAKE dataset.

#### a) Conclusion for O1 (Model Accuracy Comparison):

The results successfully met Objective 1, which was to compare the classification accuracy of the three models. The experiment showed that:

- ViT-B/16 achieved the highest accuracy.
- ResNet18 performed at a very similar level.
- MobileNetV2 produced the lowest accuracy.

This confirms that different deep learning architectures perform differently when detecting AI-generated images, and that high-capacity models outperform lightweight models in this task.

#### b) Conclusion for O2 (Best Overall Model Selection):

Objective 2, which aimed to identify the most suitable model for AI-generated artwork detection, was also achieved. Based on the results:

- ViT-B/16 was identified as the best overall model, offering the highest accuracy and most stable learning behavior.
- ResNet18 was a strong alternative, with only a very small difference in performance.
- MobileNetV2 was suitable only for low-resource environments, where speed and efficiency are more important than maximum accuracy.

These findings directly support both the research objectives and the practical goals of developing reliable AI-image detection systems.

## B. Hypothesis Evaluation

The research hypothesis stated that:

- ViT-B/16 would achieve the highest performance.
- MobileNetV2 would achieve the lowest performance.
- ResNet18 would perform between the two.

Based on the experimental results, the hypothesis is fully supported. The observed performance rankings followed the exact trend predicted in the hypothesis.

## C. Key Lessons Learned

- 1) Vision Transformers are highly effective for AI-generated image detection. The strong performance of ViT-B/16 shows that global attention and patch-based learning are powerful tools for synthetic image detection.
- 2) CNNs remain very strong baselines. Even though ViT performed best, ResNet18 achieved nearly identical results, showing that traditional CNNs are still highly useful.
- 3) Lightweight models trade accuracy for efficiency. MobileNetV2 performed much faster and required fewer resources, but had significantly lower accuracy.
- 4) Balanced datasets are critical for fair evaluation. Using a balanced CIFAKE test set helped prevent performance bias toward either class.
- 5) Multiple metrics are required for reliable validation. Accuracy alone is not enough; confusion matrices and ROC curves provided deeper insight into model behavior.

## D. Future Work

Several improvements and extensions can be made based on this study:

- 1) Testing on larger and more diverse datasets, using higher-resolution AI-generated artwork and art-specific datasets
- 2) Multi-class AI generator classification. Instead of binary REAL vs FAKE classification, future models could identify which AI model generated the image
- 3) Explainable AI integration. Tools like Grad-CAM and attention visualization can be used to better understand what visual features the models rely on.

## E. Final Conclusion

This study demonstrates that deep learning models can reliably detect AI-generated artwork. Among the three tested architectures, ViT-B/16 was the most effective, followed closely by ResNet18, while MobileNetV2 provided an efficient but less accurate alternative.

The findings of this research contribute to the growing effort to protect digital authenticity, prevent misinformation, and improve trust in visual media.

## ACKNOWLEDGMENT

We used the following tools for grammar suggestions/spelling corrections:

- Grammarly
- ChatGPT

## REFERENCES

- [1] A. Valyaeva, "AI image statistics: How much content was created by AI," Everypixel Journal, Aug. 15, 2023. [Online]. Available: <https://journal.everypixel.com/ai-image-statistics>
- [2] Interaction Design Foundation, "What is AI-generated art?," Dec. 8, 2023. [Online]. Available: <https://www.interaction-design.org/literature/topics/ai-generated-art>
- [3] Z. Thomas, "AI-assisted vs AI-generated: Is it important in 2025," Miami Entertainment & IP Lawyer, Jan. 16, 2025. [Online]. Available: [https://zthomaslaw.com/ai-assisted-vs-ai-generated/?st\\_source=ai\\_mode](https://zthomaslaw.com/ai-assisted-vs-ai-generated/?st_source=ai_mode)
- [4] J. J. Bird and A. Lotfi, "CIFAKE: Image classification and explainable identification of AI-generated synthetic images," 2023. [Online]. Available: <https://doi.org/10.48550/arxiv.2303.14126>
- [5] M. Sharafudeen, A. J., and V. Chandra S. S., "Leveraging vision attention transformers for detection of artificially synthesized dermoscopic lesion deepfakes using Derm-CGAN," \*Diagnostics\*, vol. 13, no. 5, Art. no. 825, 2023. [Online]. Available: <https://doi.org/10.3390/diagnostics13050825>
- [6] D. Park, H. Na, and D. Choi, "Performance comparison and visualization of AI-generated-image detection methods," \*IEEE Access\*, vol. 12, pp. 62609–62627, 2024. [Online]. Available: <https://doi.org/10.1109/ACCESS.2024.3394250>
- [7] F. Abbas and A. Taeihagh, "Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence," \*Expert Systems with Applications\*, vol. 252, Art. no. 124260, 2024. [Online]. Available: <https://doi.org/10.1016/j.eswa.2024.124260>
- [8] J. J. Bird, "CIFAKE: Real and AI-generated synthetic images," Kaggle. [Online]. Available: <https://www.kaggle.com/datasets/birdy654/cifake-real-and-ai-generated-synthetic-images>
- [9] S. J. D. Prince, \*Understanding Deep Learning\*. Cambridge, MA: MIT Press, 2023.
- [10] ImageNet, "ImageNet." [Online]. Available: <https://www.image-net.org/>
- [11] "ViT\_B\_16," Torchvision Main Documentation. [Online]. Available: [https://docs.pytorch.org/vision/main/models/generated/torchvision.models.vit\\_b\\_16.html](https://docs.pytorch.org/vision/main/models/generated/torchvision.models.vit_b_16.html)
- [12] "MobileNet V2," Torchvision Main Documentation. [Online]. Available: <https://docs.pytorch.org/vision/main/models/mobilenetv2.html>
- [13] "ResNet18," Torchvision Main Documentation. [Online]. Available: <https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet18.html>