Six-Channel Deep Learning Approach for Detecting AI-Generated Images

R Avantikaa[1,*,†], S K Sangeetha[1,†], B Madhuri[1,†] and A VijayaLakshmi[1,†]

[1]*Sri Sivasubramaniya Nadar College of Engineering, Chennai*

Abstract

The detection of AI-generated synthetic images is critical for ensuring content authenticity and mitigating the risks of misinformation. Our lightweight six-channel CNN model and modified ResNet6Channel is designed for efficiency, robustness across diverse generation techniques, and interpretable classification. Experiments on a test set of 10,000 samples show that REAL and FAKE images can be detected almost evenly, with an overall accuracy of 52.25 percent and an F1-score of 0.5225. They also highlight patterns of misclassification that need more investigation.

1. Introduction

The Synthetic Images Task aims to distinguish synthetic images from real ones, a challenge that has received considerable attention due to the rapid proliferation of AI-generated media [1]. This problem is particularly complex, as synthetic images exhibit a wide range of characteristics: some involve subtle pixel-level manipulations, others are fully AI-generated, and yet others employ advanced techniques that closely mimic authentic photography. Such variability poses significant challenges, as conventional image analysis methods often fail to generalize across different synthesis approaches. In this study, we focus on the binary classification subtask, developing a deep learning-based system to differentiate real from synthetic images. To enhance robustness and generalization, we incorporate data augmentation techniques.

2. Related Work

Early studies primarily focused on identifying pixel-level artifacts and inconsistencies characteristic of early-generation synthetic images [2]. Wang [2] proposed a Siamese CNN to detect GAN-generated faces by learning fine-grained spatial features. Lokner Lađević et al. [3] demonstrated that lightweight CNN architectures can efficiently detect AI-generated synthetic content.

Recent studies have explored multi-channel and augmented-input designs for image forensics. Alotaibi et al. [4] proposed a hybrid deep learning model using multiple input representations to verify the authenticity of social media posts. Our work builds on these insights by integrating RGB, edge, frequency, and noise channels into a unified six-channel input, which, to our knowledge, has not been systematically evaluated for real-world AI-generated images.

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online

*Corresponding author.

†These authors contributed equally.

avantikaa2470004@ssn.edu.in (R. Avantikaa); sangeetha2470072@ssn.edu.in (S. K. Sangeetha); madhuri2470035@ssn.edu.in (B. Madhuri); vijayalakshmia@ssn.edu.in (A. VijayaLakshmi)

3. Approach

3.1. Overview

The design prioritizes the efficiency of the model in interpreting, reproducing, and its robustness to different synthetic generation techniques. Figure 1 illustrates the core stages of our workflow: (1) data loading and pre-processing, (2) model initialization and configuration, (3) training with adaptive optimization and early stopping, and (4) threshold-based inference for classification.

3.2. Data Preparation

The dataset contains both real and AI-generated images, split into training and validation sets to evaluate model generalization. Images were resized to 224×224, normalized, and augmented with random flips and color jittering to improve robustness.

Six-Channel Input Representation: Each input image is represented by six channels:

1. Three standard RGB channels capturing color and texture,
2. One frequency-domain channel obtained from the Discrete Cosine Transform (DCT),
3. One edge-intensity map using the Sobel operator,
4. One noise residual channel extracted through high-pass filtering.

Rationale for Six Channels: The six channels were selected to capture complementary cues for distinguishing synthetic images from real ones, including color realism, structural integrity, and compression artifacts. Prior work has shown that multi-channel and frequency-domain features improve robustness against diverse generative models [4], and preliminary experiments confirmed that combining RGB, frequency, edge, and noise channels provides the best discriminative power.

Data were loaded and batched using PyTorch `DataLoader`, forming a single input tensor of shape (6, 224, 224) for both CNN6Channel and ResNet6Channel.

3.3. Model Architecture

We evaluated two architectures with the same six-channel input:

CNN6Channel: A lightweight baseline with three convolutional layers, batch normalization, ReLU activations, and max-pooling. The final fully-connected layer outputs a single logit representing the probability of an image being synthetic. This architecture isolates the effect of the six-channel input without deep residual connections.

ResNet6Channel: Adapted from ResNet-50 [5], the initial convolutional layer is modified to accept six input channels. Residual connections preserve gradient flow and enable the network to capture richer hierarchical and frequency-domain features. The final fully-connected layer outputs a single logit for binary classification.

Dual-Network Motivation and Results: CNN6Channel serves as a controlled baseline to assess the contribution of the six-channel input alone. This is expanded upon by ResNet6Channel, which uses residual learning to enhance feature representation. Preliminary experiments show CNN6Channel achieved a validation accuracy of 72%, while ResNet6Channel reached 85% under identical conditions, highlighting the added benefit of deeper architectures for this input design. Since the ResNet6Channel achieved superior validation performance, it was selected for evaluation on the official test set, and its results are reported in Section 4.

Novelty: Our approach was inspired by the multi-input text-based model of Alotaibi et al. [4], originally proposed for verifying the authenticity of social media posts. We adapt this concept to the visual domain by integrating RGB, frequency, edge, and noise channels into a

unified six-channel input, which, to our knowledge, has not been systematically evaluated on real-world AI-generated images.

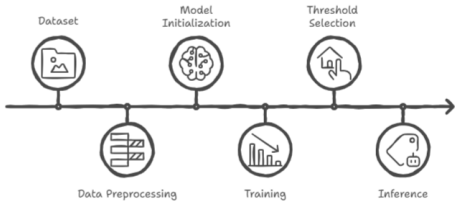

Figure 1: Overview of the model architecture and workflow.

3.4. Training Strategy

The Adam optimizer was used with an initial learning rate of 0.001, selected for its adaptive step size and efficient convergence in CNN training. A learning rate scheduler (`ReduceLROnPlateau`) was employed to automatically decrease the learning rate when validation performance plateaued, preventing stagnation. Training ran for 15 epochs (batch = 32) with checkpointing on validation AUC improvement. This ensured that the best-performing model was preserved for final evaluation and inference.

4. Results and Analysis

4.1. Quantitative Results

The evaluation of the model was conducted on a test set of 10,000 samples, equally divided between the REAL and FAKE classes. The resulting confusion matrix heatmap, shown in Figure 2, reveals that 2612 REAL instances were correctly classified, while 2388 were misclassified as FAKE. Similarly, 2613 FAKE instances were correctly predicted, with 2387 misclassified as REAL.

Table 1
Summary of performance metrics for the final model.

Metric	Accuracy	Precision	Recall	F1-score
Value	0.5225	0.5225	0.5226	0.5225

4.2. Qualitative Analysis

A detailed analysis of the confusion matrix and probability-based metrics indicates a nearly symmetric pattern of misclassification between REAL and FAKE classes, suggesting that the current features provide limited discriminatory power. Consequently, the model captures few subtle distinctions, resulting in performance close to random guessing. Threshold analysis further shows that improving F1-score via extreme thresholds leads to substantial misclassification, highlighting the trade-off between precision and recall. These findings suggest that the model requires enhanced feature engineering, inclusion of additional contextual or linguistic information, and potentially more advanced modeling approaches, such as ensemble or neural network-based methods, to better capture complex patterns in the data.

4.3. Probability-Based Performance Metrics

The probability-based evaluation reveals limited discriminative ability: the model achieves an ROC AUC of 0.5200 and an average precision of 0.5136, indicating performance only slightly above random.

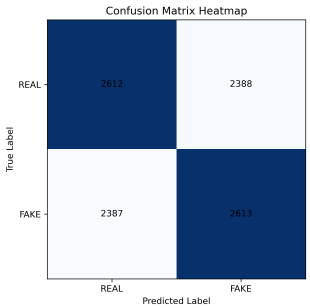

Figure 2: Confusion matrix heatmap showing the distribution of true and predicted classes.

5. Discussion and Outlook

The model demonstrates moderate performance, with misclassifications nearly balanced between REAL and FAKE classes, indicating limited discriminative power in the current features. Threshold and probability-based analyses highlight trade-offs between precision and recall, showing that optimizing a single metric does not guarantee reliable predictions. These findings suggest that future work should focus on improved feature representation, incorporation of contextual information, and exploration of more advanced modeling approaches. Overall, the evaluation highlights the model's limitations and guides directions for refinement.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] O. Papadopoulou, M. Schinas, R. Corvi, D. Karageorgiou, C. Koutlis, F. Guillaro, E. Gavves, H. Mareen, L. Verdoliva, S. Papadopoulos, Synthetic images at mediaeval 2025: Advancing detection of generative ai in real-world online images, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025.

[2] J. Wang, An eyes-based siamese neural network for the detection of gan-generated faces, Frontiers in Signal Processing 3 (2022) 918725. URL: https://www.frontiersin.org/articles/10.3389/frsip.2022.918725/full.

[3] A. L. Lađević, T. Kramberger, R. Kramberger, D. Vlahek, Detection of ai-generated synthetic images with a lightweight cnn, AI 5 (2024) 1575–1593. URL: https://doi.org/10.3390/ai5030076. doi:10.3390/ai5030076.

[4] S. Alotaibi, N. Alsubaie, O. Alsharif, Use of multiple inputs and a hybrid deep learning model for verifying the authenticity of social media posts, Electronics 14 (2025) 1184. URL: https://www.mdpi.com/2079-9292/14/6/1184. doi:10.3390/electronics14061184.

[5] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778. URL: https://doi.org/10.1109/CVPR.2016.90. doi:10.1109/CVPR.2016.90.