International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 12 | Dec - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Detection of AI-Generated Images Using CNN

Mrs. P Smitha[1], Bhoomika N[2], Bhoomika R[3] , Gowtham B R[4] , Hemanth Gowda K[5]

[1] *Assistant Professor, Dept of ISE, East West Institute of Technology, Bengaluru*
[2,3,4,5] *Student, Dept Of ISE, East West Institute of Technology, Bengaluru*

Abstract

The rapid rise of modern artificial intelligence has greatly advanced text-to-image generation, enabling the creation of synthetic visuals that are often indistinguishable from real photographs. In this study, we propose a novel detection framework built on a gated-expert Convolutional Neural Network (CNN) to identify images created by various generative models, such as BigGAN, GLIDE, VQDM, and Stable Diffusion. The proposed system employs transfer learning using ResNet-50, where each expert network is individually fine-tuned for a specific generator model. The gating module assigns dynamic importance to each expert based on the characteristics of the input image. Experimental results show that the proposed approach performs better than conventional single-CNN and basic ensemble techniques, particularly when the training data are limited. Although generator-specific models perform well in their own domains, they show a decline in accuracy for unseen generators. In contrast, our gated architecture consistently maintained an accuracy above 96% across mixed datasets. This study highlights the importance of ensemble-based solutions and diverse datasets in detecting synthetic media, offering a promising direction for future research on generalizable and adaptable detection systems.

Keywords: Artificial Intelligence, CNN, Synthetic Images, Deep Learning, Transfer Learning, Ensemble Networks

1. INTRODUCTION

Generative AI has rapidly transformed the production of digital content. Current generative models can create hyper-realistic images that can easily deceive human viewers. Although these advancements support creative and industrial applications, they also introduce risks, such as identity misuse, misinformation, and security vulnerabilities.

A major challenge in AI-generated image detection is that most existing models perform well only when tested on data from the same generator used for training. When exposed to images produced using different architectures, these models experience a significant decrease in performance. This lack of cross-generator generalization is a major obstacle in real-world environments in which images originate from various AI systems. Researchers have noted that traditional CNNs fail to adapt when encountering unfamiliar synthetic patterns when trained on a single generator.

1.1 Research Contribution

This study proposes a gated-expert CNN framework that can effectively detect synthetic images produced by multiple generative systems. By combining several specialized CNN models and controlling their influence through a gating network, the proposed architecture dynamically adjusts to the visual characteristics of each image. This approach improves robustness, reduces overfitting, and provides strong adaptability to generators that have not been previously seen.

1.2 PROBLEM STATEMENT

Most existing CNN-based detectors struggle to identify images generated by different AI models consistently. Networks trained on a single type of generator show strong results only within that domain; however, their performance drops drastically with unfamiliar generators. This limitation reduces the reliability of real-world applications, where images from various generative systems frequently coexist. Hence, there is a need for a more reliable, scalable, and generalizable detection model capable of identifying synthetic images regardless of the underlying generation method

1.3 KEY OBJECTIVES

The main objectives of this study are as follows:

1. A reliable classification system capable of differentiating real and AI-generated images from diverse generative frameworks was developed.
2. CNN-based expert models were developed using transfer learning for improved feature extraction.
3. A gated-expert mechanism that dynamically selects and weighs expert predictions is designed.
4. Evaluate performance across accuracy, precision, recall, and generalization to the unseen data.
5. Test system adaptability to unfamiliar AI-generated images.
6. The computational efficiency was analyzed to ensure suitability for practical deployment.

2. RELATED WORK AND LITERATURE SURVEY

2.1 Synthetic Image Detection Approaches

Bird and Lotfi proposed CIFAKE, a classifier designed to distinguish Stable Diffusion images from CIFAR-10 real images. Their CNN-based system achieved high accuracy but was limited to low-resolution images and a single generator.

Iqbal et al. focused on detecting deepfake facial images using VGG-19 with data augmentation. Their approach produced around 90% accuracy for face-related datasets but did not generalize well to broader categories of synthetic images

2.2 Gradient-Based and Universal Detection Methods

Tan et al. explored gradient-based features to identify subtle artifacts produced by GANs. Their hybrid CNN-Transformer model achieved good results but required heavy computation and performed poorly on images with minimal artifacts.

Ojha et al. designed a universal detector using CLIP and transformer models. Their system showed excellent cross-generator generalization, achieving approximately 98.8% accuracy, although its complexity demands high processing resources.

2.3 Benchmark Datasets and Ensemble Approaches

Zhu et al. introduced GenImage, a large dataset of real and synthetic images from multiple generators. Standard CNNs, such as ResNet-50, performed extremely well on familiar generators but showed a sharp performance decline on unseen generators, reinforcing the need for generalizable models.

Saskoro et al. proposed a gated-expert CNN similar to our approach, which showed high accuracy across multiple generators. However, their system still faced difficulties when working with completely new types of generators.

2.4 Statistical and Multi-Modal Approaches

Bonettini et al. used Benford's law to detect GAN-generated images. Although efficient and explainable, this method struggled with high-quality or adversarially trained GANs.

Singh and Sharma examined image credibility within social media using a combination of CNNs and LSTMs, demonstrating their effectiveness for misinformation detection, but they were not tailored specifically for synthetic image classification.

3. METHODOLOGY

3.1 Dataset Preparation

The dataset comprised real and synthetic images generated using BigGAN, GLIDE, VQDM, and Stable Diffusion. All images were preprocessed by resizing (299×299 pixels), normalization, and augmentation to improve generalization.

3.2 Expert CNN Models

Each expert was built using a ResNet-50 backbone pretrained on ImageNet. The models were fine-tuned separately using images from specific generators. Training utilized:

- Adam optimizer
- Binary cross-entropy loss
- Dropout and early stopping to prevent overfitting
- Adaptive batch sizes

3.3 Gating Network

The gating module determines the contribution of each expert to the overall model. It includes:

- A lightweight feature extraction layer
- Hidden layers with ReLU activation
- Softmax output to generate normalized weights

The final prediction is computed as follows:

$y\_final = \Sigma(w_i \times y_i)$

3.4 Training Procedure

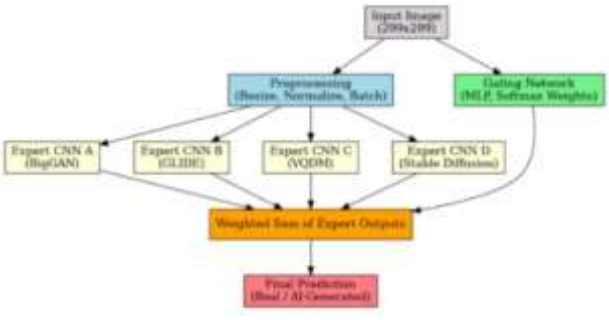
The training process consists of three steps.

1. Train individual experts on separate generator data sets.
2. The gating network is trained while keeping the expert weights frozen.
3. Optional fine-tuning is performed to improve collaboration between components.

3.5 System Architecture

The complete system includes the following:

- Expert Models: Modified ResNet-50 networks trained on specific generators
- Gating Network: Multi-layer perceptron
- Aggregation Module: Weighted output combination producing final classification



4. RESULTS AND DISCUSSION

4.1 Performance Overview

The proposed gated-expert system consistently outperforms individual CNNs and standard ensembles.

Model Configuration	Accuracy	Precision	Recall	F1-Score
Individual Expert (BigGAN)	99.1%	99.3%	98.9%	99.1%
Individual Expert (GLIDE)	98.7%	98.9%	98.5%	98.7%
Individual Expert (VQDM)	99.4%	99.5%	99.3%	99.4%
Individual Expert (Stable Diffusion)	98.2%	98.5%	97.9%	98.2%
Averaging Ensemble	94.3%	94.7%	93.9%	94.3%
Voting Ensemble	95.1%	95.4%	94.8%	95.1%
Unified Model	95.8%	96.1%	95.5%	95.8%
Gated Expert Model	**96.4%**	**96.8%**	**96.1%**	**96.4%**

4.2 Cross-Generator Performance

- Within-generator testing: Experts achieved an accuracy of above 98 %.
- For unfamiliar generators, the accuracy decreased to 65–75%.
- Gated architecture: Maintained accuracy above 94% for all generator types.

4.3 Computational Efficiency

- Reduced training duration by 40% compared to unified models
- 30% lower memory usage during inference
- Classification time under 2 seconds per image on typical GPUs

4.4 Ablation Findings

- Removing the gating network decreased accuracy by about 4.2%
- Using a unified model instead of specialized experts reduced performance
- Removing transfer learning increased training time by nearly threefold

5. CONCLUSIONS

This study presents a gated-expert CNN architecture that can detect AI-generated images from several major generative models. Unlike traditional CNNs that suffer from limited generalization, the proposed design combines specialized expert networks with a dynamic gating mechanism, allowing the system to adapt to diverse input types. The model achieved an accuracy of above 96% and offered strong generalization, computational efficiency, and adaptability. Although the performance decreases slightly for completely novel generators, the framework remains a strong candidate for real-world applications in digital forensics, authentication and security.

ACKNOWLEDGEMENT

We extend our gratitude to the Department of Computer Science and Engineering for supporting this work with guidance and the computational resources. We also acknowledge the research community for providing the open-source datasets and tools used in this study.

REFERENCES

[1] Bird, J.J., and Lotfi, A. (2023). "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images." arXiv preprint arXiv:2303.14126.

[2] Iqbal, F.; Abbasi, A.; Javed, A. R. (2022). "Data Augmentation-Based Novel Deep Learning Method for Deepfaked Images Detection." ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 18, no. 3, pp. 1-15.

[3] Tan, C., Zhao, Y., Wei, S., Gu, G., and Wei, Y. (2023). "Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12105-12114.

[4] Ojha, U., Li, Y., and Lee, Y.J. (2023). "Towards Universal Fake Image Detectors That Generalize Across Generative Models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 24480-24489.

[5] Zhu, M., et al. (2023). "GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image." arXiv preprint arXiv:2306.08571.

[6] Saskoro R. A. F., Yudistira, ; Fatyanosa, T.N. (2024). "Detection of AI-Generated Images From Various Generators Using Gated Expert Convolutional Neural Network." IEEE Access, vol. 12, pp. 147772-147783, doi: 10.1109/ACCESS.2024.3466614.

[7] Patel, M., Fatangare, S., Nasare, A., and Pachpute, A. (2022). "Image-Dev: An Advance Text to Image AI Model." Proceedings of IEEE PuneCon, pp. 1-6.

[8] Bonettini, N., Bestagini, P., Milani, S., and Tubaro, S. (2021). "On the Use of Benford's Law to Detect GAN-Generated Images." Proceedings of the 25th International Conference on Pattern Recognition (ICPR), pp. 5495-5502.

[9] Mohammed, A., and Kora, R. (2023). "A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges." Journal of King Saud University - Computer and Information Sciences 35, no. 2, pp. 757-774.

[10] Singh; Sharma, D.K. (2022). "Predicting Image Credibility in Fake News Over Social Media Using Multi-Modal Approach." Neural Comput. Appl. 34, no. 24, pp. 21503-21517.