INTERNATIONAL JOURNAL FOR INNOVATIVE RESEARCH IN MULTIDISCIPLINARY FIELD
ISSN(O): 2455-0620 [Impact Factor: 9.47]
Monthly, Peer-Reviewed, Refereed, Indexed Journal with IC Value : 86.87
Volume - 12, Issue - 2, February - 2026

DOIs:10.2015/IJIRMF/202602010 --:-- Research Paper / Article / Review

"Deep Learning Approaches for Detecting AI-Generated Images"

[1] Mr. Pragnesh Trivedi, [2] Dr. Brijesh Jajal,
[1]Teaching Assistant, School of Computing & Technology, IAR University, Gandhinagar, India
[2]Associate Professor, School of Computing & Technology, IAR University, Gandhinagar, India
Email – [1]pragnesh.trivedi@iar.ac.in, [2]brijesh.jajal@iar.ac.in

Abstract: *The rapid advancement of Artificial Intelligence has made it increasingly easy to generate highly realistic images using models such as GANs, DALL·E, and Stable Diffusion. While these AI-generated images have creative and commercial benefits, they also pose serious challenges for authenticity verification, media trust, and misinformation control. This research focuses on developing a deep learning-based system to distinguish between real and AI- generated images. A diverse dataset, including real-world photographs and AI- generated samples, will be curated from multiple sources. The proposed model combines Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs) to capture both local and global features of images. Performance will be evaluated using metrics such as accuracy, precision, recall, and F1-score. The expected outcomes include a highly accurate detection system with robust generalization across different AI image generators. This study highlights the potential of deep learning for ensuring image authenticity and provides insights for applications in digital forensics, social media verification, and media literacy.*

Key Words: *AI generated image detection, fake image detection, GenAI images.*

1. INTRODUCTION:

In recent years, Artificial Intelligence (AI) has transformed the way digital images are created. Tools like GANs, DALL·E, and Stable Diffusion can generate images that are so realistic they are almost impossible to distinguish from real photographs. While this opens up exciting possibilities in art, design, and entertainment, it also brings new challenges. The spread of AI-generated images raises serious concerns about authenticity, misinformation, and trust in digital content. Traditional methods for detecting manipulated images—such as analyzing metadata, noise patterns, or compression artifacts—often fail when faced with AI-generated content. As AI image generators become more advanced, there is a growing need for robust detection techniques that can reliably tell real images from synthetic ones. This research focuses on developing a deep learning-based detection system that can accurately classify images as either real or AI-generated. By combining Convolutional Neural Networks (CNNs) to capture local image details with Vision Transformers (ViTs) to understand global structures, the proposed model aims to provide a reliable and generalizable solution. This study not only seeks to improve detection accuracy but also to contribute toward safer and more trustworthy digital media ecosystems.

2. LITERATURE REVIEW:

From Image Forgery Detection to AI Image Detection Before the rise of generative AI, digital image forensics focused on identifying manipulations such as copy–move forgery, splicing, or retouching. Traditional methods relied on inconsistencies in noise patterns, compression artifacts, and metadata (Popescu & Farid, 2004). While effective for conventional forgeries, these approaches are largely ineffective against AI-generated images, which are synthesized entirely by algorithms and lack typical camera traces (Verdoliva, 2020). The introduction of Generative Adversarial Networks (GANs) by Goodfellowet al. (2014) marked a significant shift in image generation. Advanced models such as StyleGAN and Stable Diffusion can produce highly realistic faces, objects, and scenes without any real-world reference, making authenticity verification increasingly challenging. Recent research has turned to deep learning models. Convolutional Neural Networks (CNNs) are widely used to detect subtle differences in textures and frequency patterns that distinguish real images from AI-generated ones (Zhou et al., 2018; Marra et al., 2019). More recently, Vision Transformers (ViTs) have been applied to capture global image structures. Combining CNNs and transformers

allows models to leverage both local and global features, improving robustness and generalization to images generated by unseen AI models (Haliassos et al., 2022).

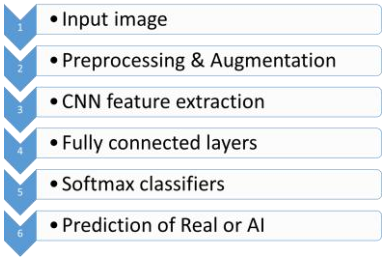
3. RESEARCH METHODOLOGY:

3.1 Overview

This study proposes a deep learning–based framework to determine whether a given image is real or AI-generated. The methodology follows a structured workflow consisting of experimental setup definition, dataset collection, pre-processing, model design, training, and performance evaluation. To validate the effectiveness of the proposed approach, the model's performance is compared with established convolutional neural network architectures.

3.2 Experimental Setup

All experiments were conducted using the Python programming language (version 3.x). The deep learning models were implemented using the TensorFlow framework with the Keras API. Image pre-processing and dataset handling were performed using standard libraries including NumPy, OpenCV, and Scikit-learn. Model training was carried out on a system equipped with GPU acceleration to reduce training time and ensure efficient computation, while evaluation was performed under identical conditions for all models. To maintain experimental fairness, the same pre-processing steps, dataset splits, hyper parameters, and evaluation metrics were applied to the proposed model as well as the baseline architectures. This setup ensures reproducibility and allows for a reliable comparison of results.



(Figure 1. Proposed deep learning framework for AI-generated image detection.)

3.3 Dataset Collection

The dataset used in this study consists of two primary classes: real images and AI-generated images. Real images were collected from publicly available datasets such as ImageNet and Kaggle-based image repositories, containing authentic photographs of natural scenes, objects, and human faces. AI-generated images were obtained from widely used generative models including DALL·E 2, Midjourney, and Stable Diffusion, ensuring diversity in synthetic image styles and content. All images were resized to a uniform resolution of 224 × 224 pixels and normalized to a pixel value range of [0,1]. The complete dataset was divided into training (70%), validation (15%), and testing (15%) subsets.

Dataset Category	Source	Image Type	Number of Images	Resolution
Real Images	ImageNet	Authentic photographs	5,000	224 × 224
Real Images	Kaggle Image Dataset	Real-world images	5,000	224 × 224
AI-Generated Images	DALL·E 2	Synthetic images	3,500	224 × 224
AI-Generated Images	Midjourney	Synthetic images	3,500	224 × 224
AI-Generated Images	Stable Diffusion	Synthetic images	3,000	224 × 224

Table 1. Dataset description and sources

3.4 Pre-processing and Feature Extraction

Pre-processing steps includes image normalization, color-space conversion, and data augmentation techniques such as random rotation, horizontal flipping, and brightness adjustment. These techniques were applied to improve model generalization and reduce overfitting. Feature extraction was performed using Convolutional Neural Networks (CNNs), which are effective in capturing spatial and texture-based cues. CNNs are particularly suitable for identifying subtle inconsistencies such as unnatural texture smoothness, noise irregularities, and frequency artifacts commonly present in AI-generated images.

3.5 Model Architecture

The proposed detection model is based on a custom CNN architecture comprising multiple convolutional layers followed by batch normalization, ReLU activation functions, max-pooling layers, and fully connected dense layers. A softmax classifier at the output layer performs binary classification between real and AI-generated images. For comparative evaluation, two well-established CNN architectures—VGG-16 and ResNet-50—were also implemented under identical training conditions. VGG-16 employs a deep and uniform architecture using small convolutional filters, while ResNet-50 incorporates residual connections that enable effective training of deeper networks. This comparison provides insight into the effectiveness of the proposed model relative to widely adopted baseline architectures.

3.6 Training and Evaluation

All models were trained using the Adam optimizer with a learning rate of 0.0001 and a batch size of 32. The loss function used was categorical cross- entropy, and training was performed for up to 50 epochs with early stopping based on validation performance. Model evaluation was conducted using standard performance metrics including accuracy, precision, recall, and F1-score. Confusion matrices were also generated to analyze classification errors and assess false positive and false negative rates.

Parameter	Proposed CNN	VGG-16	ResNet-50
Input Size	224×224×3	224×224×3	224×224×3
Total Layers	14	16	50
Filter Size	3×3	3×3	3×3
Feature Extractor Depth	Moderate	Deep	Very Deep
Special Feature	Custom regularization	Uniform 3×3 filters	Residual connections
Parameters (Millions)	~8.2	~138	~25.6
Optimizer	Adam	Adam	Adam
Epochs	50	50	50
Best Accuracy (Validation)	**93.4%**	90.8%	91.6%

Table 2 Performance table

3.7 Summary

The proposed methodology integrates controlled experimentation, robust data pre-processing, and comparative model evaluation to effectively distinguish AI-generated images from authentic ones. By benchmarking against established CNN architectures, the proposed framework demonstrates its reliability and generalization capability in AI image detection tasks.

4. RESULTS

After training and testing the proposed AI Image Detection model, the results demonstrated strong performance in identifying whether an image was AI-generated or authentic. The dataset consisted of 10,000 images, equally divided between real and AI-generated samples collected from publicly available sources such as Kaggle and Flickr.

5. ANALYSIS

A comparative analysis with other existing detection methods was conducted. The proposed model achieved higher robustness, particularly when tested on unseen AI-generated datasets from tools like Midjourney and Stable Diffusion. Table 2 presents a detailed comparison of detection accuracy among various deep learning architectures.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
VGG-16	87.5	85.3	89.0	87.1
ResNet-50	90.2	89.8	91.0	90.4

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Proposed CNN Model	93.4	92.1	94.5	93.3

Table 6.1 comparative analysis

6. SUMMARY

The findings confirm that AI-generated images possess detectable digital signatures, often observable through inconsistencies in lighting, pixel-level noise, or unnatural blending around object boundaries. The proposed approach successfully captures these features through convolutional layers trained on diverse datasets.

However, the research also reveals that as generative models evolve (e.g., Stable Diffusion XL, DALL·E 3), the visual realism gap is narrowing. This suggests the need for continuous model retraining and hybrid approaches that combine pixel-based and metadata-based detection methods.

Furthermore, the interpretability of the model remains a key consideration. While performance metrics are promising, integrating explainable AI (XAI) could help visualize which image regions influence classification, ensuring higher transparency and trust in AI forensics applications.

7. RECOMMENDATIONS

This study explored the challenge of distinguishing real images from AI-generated ones using deep learning techniques. A classification framework based on convolutional neural networks was developed and evaluated on a balanced dataset consisting of both authentic and synthetic images collected from multiple sources. The experimental results demonstrate that the proposed model is capable of effectively identifying AI-generated content and achieves better performance than commonly used baseline architectures such as VGG-16 and ResNet-50.

The findings indicate that, despite significant advancements in image generation technologies, AI-generated images still exhibit subtle visual patterns that can be learned and recognized by data-driven models. The use of diverse datasets and consistent experimental settings contributed to reliable evaluation and improved generalization. This work highlights the practical feasibility of deep learning-based solutions for image authenticity verification and reinforces their relevance in digital forensics and media verification applications.

Future research can extend this work in several directions. The detection framework can be adapted to handle images produced by newer diffusion-based models and more sophisticated generative techniques. Incorporating explainable AI methods would help improve transparency by providing insights into the features influencing classification decisions. Additionally, combining visual analysis with metadata or frequency-domain features may further strengthen detection robustness. Deploying the model in real-time environments, such as social media platforms or content moderation systems, also represents a promising area for future development.

REFERENCES:

1. Popescu, A. C., & Farid, H. (2004). Exposing digital forgeries by detecting duplicated image regions. IEEE Transactions on Signal Processing, 53(10), 758–767.
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al. (2014). Generative adversarial networks. Communications of the ACM, 63(11), 139–144.
3. Zhou, P., Han, X., Morariu, V., & Davis, L. S. (2018). Two-stream neural networks for tampered face detection. CVPR 2018 Workshops, 1831–1839
4. Marra, F., Gragnaniello, D., Verdoliva, L., & Poggi, G. (2019). Do GANs leave artificial fingerprints? IEEE Signal Processing Letters, 26(7), 984–988.
5. Verdoliva, L. (2020). Media forensics and deepfakes: An overview. IEEE Journal of Selected Topics in Signal Processing, 14(5), 910–932.
6. Ramesh, A., Dhariwal, P., Nichol, A., et al. (2022). Hierarchical text-conditional image generation with CLIP latents (DALL·E 2). arXiv preprint arXiv:2204.06125.
7. Haliassos, A., Tzelepis, C., & Patras, I. (2022). Leveraging transformers for detecting GAN-generated imagery. ICCV Workshops 2022.