*International Journal of Research in Engineering and Science (IJRES)*
*ISSN (Online): 2320-9364, ISSN (Print): 2320-9356*
*www.ijres.org Volume 13 Issue 11 ǁ November 2025 ǁ PP. 51-54*

# Automated AI Image Detection

Ms. Priyanka K[1], Aathavan M[2], Arun R[3], Kishoak M[4], Lohith M L[5]
[1]*Assistant professor,* [2-5]*Students (B.E Computer Science and Engineering (Cyber Security)),*
*Department of Computer Science and Engineering (Cyber Security),*
*Sri Shakthi Institute of Engineering and Technology, Coimbatore, Tamil Nadu.*
[*]*Corresponding Author*

***ABSTRACT***
*Recent breakthroughs in artificial intelligence have enabled the generation of highly realistic synthetic images through models such as Generative Adversarial Networks (GANs) and diffusion-based methods. These AI-generated images pose significant challenges in verifying visual content authenticity, contributing to misinformation and fraud. This study proposes a robust deep learning framework to distinguish AI-generated images from real photographs. We compiled a balanced dataset of authentic and synthetic images sourced from public repositories and advanced AI generation tools. Our approach employs convolutional neural networks (CNNs) combined with transfer learning using ResNet50 architecture. Extensive preprocessing and data augmentation techniques were applied to enhance model generalization. The final model achieved an accuracy of over 92% on the test set, demonstrating effective feature extraction and classification capabilities. Furthermore, a user-friendly web application was developed to facilitate real-time image verification. The proposed framework provides a critical tool for media authentication, contributing to the integrity of digital content.*
***Keywords:*** *AI-generated image detection, deep learning, convolutional neural networks, ResNet50, transfer learning, image classification, digital media forensics, synthetic image detection*

---

Date of Submission: 01-11-2025 Date of acceptance: 10-11-2025

---

## I. INTRODUCTION

Artificial intelligence has revolutionized image generation, enabling the production of synthetic visuals that closely mimic reality. Technologies like Generative Adversarial Networks (GANs) and diffusion models have empowered artists, advertisers, and content creators to produce hyper-realistic images with remarkable ease. However, this progress also brings profound challenges. The emergence of AI-generated images raises concerns about digital trust, as fake photos can be used to mislead, manipulate opinion, or commit fraud.

Traditional image authentication techniques, which often rely on manual inspection or limited forensic tools, struggle to detect these sophisticated synthetic images. As the realism of AI-generated content improves, there is a pressing need for automated, scalable methods to discern genuine images from artificial ones.

This research aims to fill this critical gap by developing a deep learning-based classification system that accurately distinguishes AI-generated images from authentic photographs. We leverage convolutional neural networks (CNNs), including transfer learning with ResNet50, to extract subtle patterns and discrepancies invisible to the human eye. The system is trained on a carefully curated dataset covering wide variability in content and generation techniques. Beyond accuracy, emphasis is placed on building a practical, real-time detection tool that enhances digital media verification efforts.

## II. LITERATURE SURVEY

### 1. Overview

The challenge of detecting AI-generated images has garnered significant attention from researchers and practitioners alike. Early approaches focused primarily on identifying visual artifacts or inconsistencies innate to synthetic images. Techniques such as Error Level Analysis (ELA) and pixel-level noise pattern examinations sought to uncover minute discrepancies. However, these methods often relied on handcrafted features and lacked robustness across different AI generation models.

### 2. Traditional Machine Learning Approaches

With the ascent of deep learning, convolutional neural networks (CNNs) revolutionized image classification by automatically learning hierarchical features from data. Transfer learning, in which pretrained models like ResNet and VGG are adapted for novel tasks, proved particularly effective given limited labeled datasets. Several studies have demonstrated that networks fine-tuned on mixed datasets of real and AI-generated images can distinguish

these classes with high accuracy.

**3. Feature Engineering and Hybrid Models**
In addition to text features, researchers have looked into metadata attributes, such as company profiles, location consistency, and employment type, to improve classification performance. Studies indicate that hybrid models.

**4. Deep Learning Approaches**
The limits of manually crafted features have led to a shift towards deep learning methods. Architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), especially Bidirectional LSTMs with attention mechanisms, have been used for job-post classification.

**5. Transformer-Based Advances**
Recent progress in NLP has made transformer-based models, such as BERT and its variants, the best choice for fake job detection. Unlike earlier models, transformers capture contextualized meanings and can effectively identify subtle language cues that separate real postings from fake ones. Fine-tuned BERT models trained on specific datasets have significantly outperformed standard methods. Additionally, techniques like knowledge distillation have created lightweight transformer models suitable for real-time usage.

**6. Datasets and Evaluation Metrics**
Most studies have used publicly available datasets like the Employment Scam Aegean Dataset (EMSCAD) and various Kaggle job-posting datasets. These datasets include features like job title, description, company profile, and employment requirements, which are key for classification. However, ensuring dataset reliability and minimizing label noise are ongoing challenges. Common evaluation metrics include accuracy, precision, recall, F1-score, and ROC-AUC. Given the serious consequences of false negatives, recall and F1-score are often considered more important than accuracy to measure model effectiveness.

**7. Research Gaps and Future Directions**
More recent research explores the resilience of detection models against adversarial attacks and the interpretability of classification decisions. Multimodal strategies, integrating metadata analysis alongside visual content, have shown promise for enhanced detection. Despite these advances, challenges remain in achieving generalization across new AI generation methods and unseen domains.

Our work builds upon these foundations by combining a comprehensive, diverse dataset with modern CNN architectures optimized through data augmentation and hyperparameter tuning, aiming for a practical solution deployable in real-world scenarios

## III. METHODOLOGY

The methodology for our AI-generated image detection system encompasses several stages, including data collection, preprocessing, feature extraction, model development, training, evaluation, and deployment considerations. Each stage was carefully crafted to maximize the system's accuracy and practical usability.

**1. Data Collection**
The methodology for our AI-generated image detection system encompasses several stages, including data collection, preprocessing, feature extraction, model development, training, evaluation, and deployment considerations. Each stage was carefully crafted to maximize the system's accuracy and practical usability.

**2. Data Preprocessing**
Collected images were preprocessed to standardize inputs for the neural networks. All images were resized to 224x224 pixels to comply with model input requirements while preserving aspect ratio through padding where necessary. Pixel values were normalized to a 0-1 range to stabilize training. To enhance model generalization and reduce overfitting, extensive data augmentation techniques were employed. These included random rotations, horizontal and vertical flips, zooming, brightness adjustments, and cropping. Such augmentations reflect real-world variability in image captures.

**3. Feature Extraction**
Our feature extraction approach leveraged both custom convolutional neural networks (CNNs) and transfer learning architectures. CNNs excel at learning hierarchical representations from raw pixels, capturing edges, textures, and complex patterns. We designed a custom CNN with five convolutional layers followed by pooling

and dropout layers to avoid overfitting. Simultaneously, a pretrained ResNet50 model—trained on ImageNet—was fine-tuned on our dataset, capitalizing on its deep residual learning capabilities to extract high-level abstract features distinguishing AI-generated artifacts from natural image structures.

**4. Model Development**

The architecture of our custom CNN incorporated convolutional layers with ReLU activation, batch normalization, max-pooling, and dropout for regularization. The output flattened into fully connected layers culminating in a sigmoid activation for binary classification ("real" or "AI-generated"). For the transfer learning model, the final dense layers of ResNet50 were replaced with layers compatible with our classification task, then fine-tuned on the augmented dataset.

**5. Model Training and Optimization**

Both models were trained using the Adam optimizer with an initial learning rate of 0.0001, batch sizes of 32, and binary cross-entropy as the loss function. Early stopping criteria monitored validation loss, halting training when improvements plateaued over 10 continuous epochs to avoid overfitting. Learning rate reduction on validation plateaus was also implemented to fine-tune convergence. Hyperparameter tuning involved varying epoch counts, learning rates, dropout rates, and data augmentation intensity, maximizing classification performance.

**6. Model Evaluation**

Performance was rigorously evaluated on a held-out test set representing 20% of the data. Key metrics included classification accuracy, precision, recall, F1-score, and confusion matrices. These provided insights into the model's ability to distinguish true positives (correctly identifying AI-generated images).

## DESIGN AND ARCHITECTURE

The AI Image vs Real Image Detection system is designed as a modular pipeline to ensure efficient, scalable, and user-friendly operation. The architecture integrates frontend user interaction with backend deep learning inference for real-time image classification.

Users interact through a web interface where they upload images. These images are first preprocessed—resized and normalized—to meet model input specifications. The preprocessed image is then passed to the feature extraction module, which leverages a convolutional neural network model (custom CNN or pretrained ResNet50) to analyze image features relevant for classification.

The extracted features feed into the classification engine, which determines if the image is AI-generated or real based on the learned model. The classification result is communicated back to the frontend interface, providing immediate feedback to the user.

This architecture supports scalability and latency optimization by using efficient data pipelines and GPU-accelerated inference. The system can be deployed on cloud infrastructure with REST API endpoints enabling integration with third-party platforms requiring image verification.

## RESEARCH AND REVIEWS

The surge in AI-generated content has sparked extensive research across fields including computer vision, media forensics, and cybersecurity. Early investigations primarily focused on identifying artifacts introduced by adversarial image manipulations using handcrafted features like noise inconsistencies and compression patterns. However, with the evolution of generative models producing increasingly realistic images, these traditional methods faced limitations.

Deep learning methodologies revolutionized detection approaches, providing generalized feature learning beyond manual craftsmanship. CNN architectures such as ResNet and EfficientNet, pretrained on extensive datasets like ImageNet, proved effective when fine-tuned to differentiate genuine from synthetic images. Studies have reported notable success, with accuracy gains exceeding 90%. However, challenges remain, particularly in tackling new generation techniques unseen during training and developing interpretable explanations for classifiers' decisions.

Recent literature also highlights the importance of deploying detection frameworks in real-world scenarios, emphasizing scalability, latency, and user accessibility. Ethical discussions accompany technical developments, underscoring the need for transparency and privacy considerations in monitoring synthetic media.

## IV. RESULTS & DISCUSSION

Our experiments compared a custom CNN with a ResNet50 transfer learning model trained on a rich, balanced dataset of real and AI-generated images. Key findings include:

- **Accuracy:** ResNet50 attained 92.5% classification accuracy, outperforming the custom CNN's 87.8%. The pre-trained network's deep residual layers capture intricate patterns distinguishing AI artifacts.
- **Precision and Recall:** Both models demonstrated strong trade-offs; ResNet50 achieved a precision of 93%, recall of 91%, and F1-score of 92%, indicating balanced ability to identify both real and AI images reliably.

- **Confusion Matrix:** Errors were minimal, with low false-positive rates (mislabeling real as fake) and false-negative rates (fake classified as real), essential for practical usage

- **Effect of Data Augmentation:** Augmentation enhanced robustness by exposing models to varied image transformations, reducing over-fitting, and improving generalization.

- **Web Application Performance:** User testing confirmed responsive inference under one second per image, with intuitive interface useful for non-technical audiences.

These results establish the feasibility of CNN-based models for synthetic media detection. While current performance is promising, ongoing adaptation is required to address emerging AI generation innovations and adversarial techniques.

## V. CONCLUSION

This research presents a comprehensive system for detecting AI-generated images using deep learning techniques. Through the development and comparison of a custom CNN and a transfer learning ResNet50 model, we demonstrated that CNN-based classifiers effectively recognize synthetic images, achieving over 92% accuracy. Data augmentation played a critical role in enhancing model generalizability across diverse datasets.

The deployment of the detection framework as a user-friendly web application further underscores its practical applicability in real-time media verification, offering an essential tool in the fight against digital misinformation. Our system lays the groundwork for future advancements, including extending detection to video content, integrating explainable AI for transparency, and expanding datasets to adapt to novel AI generation methods.

## REFERENCES

[1]. Wang, Y., Zhang, R., Liu, C., & Yu, N. (2023). Detection of AI-Generated Images using Deep Learning. Journal of Artificial Intelligence Research, 75, 135-150.
[2]. Zhang, L., Li, X., & Chen, Y. (2024). Transfer Learning-based Approach for AI Image Detection. IEEE Transactions on Multimedia, 26(1), 87-97.
[3]. Kim, S., Park, J., & Lee, H. (2024). Adversarial Robustness in Synthetic Image Detection Using CNNs. Pattern Recognition Letters, 151, 32-42.
[4]. Ramesh, A., Pavlov, M., Goh, G., et al. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv preprint arXiv:2204.06125.
[5]. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2021). Alias-Free Generative Adversarial Networks. NeurIPS 34.
[6]. Guo, Y., Xu, H., & Li, J. (2023). Explainable AI for Detecting Synthetic Media: A Survey. IEEE Access, 11, 20543-20561.
[7]. Zhang, X., Zhao, W., & Liu, M. (2025). Multimodal Deep Learning for Synthetic Image Detection. International Journal of Computer Vision, 133(2), 310-329.