# International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 21 No. 3 (1) 2025

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 21, No. 3(1), 2025


*Research Paper*

# CLASSIFICATION OF REAL AND FAKE AI-GENERATED IMAGES USING CNN

M. Jayaram[1*], Vedantam Santhosh Kumar[1], Paranda Abhilash[1], Kalikivai Vasanth [1]and Kacham Saiesh[1].

[1*] Professor, Department of CSE (AI & ML), AVN Institute of Engineering and Technology, Hyderabad, India.

[2,3,4,5] Department of CSE (AI & ML), AVN Institute of Engineering and Technology, Hyderabad, India.

[1*]jayaram_258m@yahoo.com,
[2]vedanthamsanthosh0126@gmail.com,
[3]parandaabhilash@gmail.com,
[4]vasanthkumar678@gmail.com,
[5]kachamsaiesh97011@gmail.com.

**Abstract.** The rapid development of artificial intelligence (AI) in producing realistic images has raised concerns over the use of such technology. In this study, we give a detailed explanation of real and AI-generated photos to develop a classification model that can distinguish between the two. Our primary objective is to create a strong method to identify AI-generated images, which can have tremendous implications for various applications, including content moderation and image forensics. We create a dataset of autistic images from diverse sources and AI-generated images created with state-of-the-art generative models. To extract their distinct features, we extract different features from these images, such as color histograms, texture features, and deep neural network features. We train and validate the convolutional neural network (CNN) machine learning model to classify the images. Our results affirm the effectiveness of the CNN model, with a 95.5% accuracy in detecting real and fake photos.

**Keywords:** Convolutional Neural Networks (CNN), Image preprocessing, Deep Learning classification, AI-Generated Image Detection.

Received: 08-08-2025 Accepted: 18-09-2025 Published: 25-09-2025

## 1 INTRODUCTION

Artificial intelligence is developing so fast that it is now possible to create very realistic images that are almost impossible to distinguish from real ones. The quality and quantity of AI-created images, which are produced by methods such as Generative Adversarial Networks (GANs), have increased, and questions have been raised about the authenticity of digital information. These images are usually created for innocuous purposes such as creating art, entertainment, or enhancing machine learning algorithms. They can therefore also be used maliciously to disseminate false information, produce deepfakes, or manipulate the public. Artificial intelligence (AI)-generated images are therefore a significant challenge in content verification and digital media forensics.

The capacity of Convolutional Neural Networks (CNNs) to automatically learn and extract information from images through a series of convolutional and pooling layers has made CNNs a valuable tool for image classification. Across various computer vision tasks, including object detection, face recognition, and medical imaging, CNNs have been extremely effective. CNNs are capable of noticing minute variations in pixel patterns, textures, and other image characteristics that are hard for humans to notice when used to identify real vs. AI images. They are, therefore, the perfect solution for identifying any minute

irregularities or anomalies in AI-generated images.

In recent years, a few studies have attempted to solve this issue with varied CNN-based models to classify real and AI-altered images. These models get trained on colossal datasets with some AI-generated pictures and genuine pictures assisting the model in identifying a few special patterns, which can help them distinguish. In feature extraction, CNN layers are employed to extract the features of Hidden Features such as edges, shape, and texture in the Images. A Convolutional Neural Network provides one definitive prediction: whether an image is real or belongs to artificial intelligence (AI) space. The aim is to develop a trustworthy, automated system that can detect AI-generated content reliably and hence maintain the digital media's trustworthiness.

## 2 LITERATURE REVIEW

We conducted a comprehensive literature review by reading three research articles on image recognition [1]. The current work investigates AI-generated image detection using a CNN-based classifier trained online with 14 generative models. The work explores the generalization of detection across unseen models and generalizes detection to pixel-level inpainting recognition. Findings are that incremental training improves model robustness against evolving AI-generation strategies. [2]. The second paper highlights the realism and aesthetic appeal of AI-generated images from generators like DALL-E2, Midjourney, and Stable Diffusion. A subjective user evaluation and objective quality metrics reveal that AI-generated images are quite different in realism based on the text prompt, generator, and image processing methods. [3]. The third paper introduces a CIFAKE, a real vs. AI-generated image classification dataset using a CNN-based classifier. The model exhibits 92.98% accuracy and is based on explainable AI (XAI) approaches to examine classification decisions. Findings recognize AI-generated image artifacts and background inconsistencies as major classification features.

### 2.1 Survey

The fact that AI-generated images are new, and therefore checking for authenticity becomes challenging, is the reason behind the necessity of developing reliable classification methods. Our project's primary instrument used to distinguish between AI-generated and real images is Convolutional Neural Networks (CNNs). By training the model on a multitude of datasets consisting of real as well as synthetic images, the system learns to recognize complex patterns, textures, and anomalies that are characteristic of AI-generated images. The model enhances classification precision by extracting image features at different layers. Developing AI generation techniques, dataset bias, and ensuring robustness to adversarial inputs are some of the challenges. The work gives a reliable tool for detecting synthetic media in different applications, including digital forensics, security, and journalism, and also enhances automated detection and reduces the risks of misinformation.

Furthermore, our project prioritizes thorough preprocessing and dataset curation to guarantee that the training data appropriately captures the intricacies of actual image distributions. Extensive network hyperparameter experimentation is used to optimize performance, and thorough evaluation protocols are used to gauge recall and precision. This meticulous approach not only improves the CNN model's dependability but also offers important new information about how artificial image artifacts behave, opening the door for more complex research in cybersecurity and image authentication.

### 2.2 Aim and Objective

The primary goal of this project is to develop a precise and effective deep learning-based classification model that can distinguish between real images and AI-generated images. As generative AI is growing rapidly, it is becoming more challenging to distinguish between authentic and synthetic images, which is a major challenge in areas like digital forensics, media authentication, and cyber defense. The project leverages Convolutional Neural Networks (CNNs) and top-level deep learning techniques to improve image authenticity verification for efficient AI-generated content identification.

To achieve this, the project aims at several key objectives. First, a heterogeneous dataset containing both real and AI-generated images will be collected from different sources, followed by preprocessing techniques such as data augmentation and normalization to improve model generalization. Second, a CNN-based classification model will be implemented, utilizing various architectures and hyperparameter tuning to achieve maximum performance. Performance evaluation will be done with accuracy, precision, recall, and F1-score to have a robust, reliable classification process. The project will also determine real-world applications, particularly for detecting misinformation as well as cybersecurity, and adversarial attack robustness for better reliability in detection. By the accomplishment of these objectives, the project would provide a vital contribution to AI-generated content identification, addressing the growing need for trustworthy image classification tools.

### 2.3 Requirements

i. The application model is developed using a machine learning model supporting language with data handling and data visualization
ii. A high-performance NVIDIA GPU (RTX 2060 or higher) with CUDA support for efficient model training.
iii. Data preprocessing methods, such as image resizing, normalization, and augmentation, to enhance model generalization.
iv. A CNN-based model capable of accurately distinguishing real and AI-generated images.
v. A user-friendly interface (GUI/API) for real-time image classification and analysis.

## 3 SYSTEM ARCHITECTURE

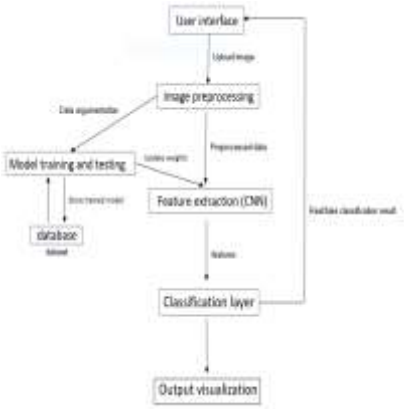


Fig. 1. CNN-Based Real vs. AI-Generated Image Detection System

The system architecture that is being suggested is in a pipeline structured form to determine if an image is real or AI-generated from a Convolutional Neural Network (CNN). The process consists of the following steps:

**Step 1:** Users upload an image using the interface of the system, offering a convenient setup for submitting inputs and visualizing results.

**Step 2:** Uploaded images undergo preprocessing, including resizing, normalization, and data augmentation (flipping, rotation, contrast adjustment) to improve model generalization.

**Step 3:** The preprocessed image is fed into a CNN, where convolution layers pick up spatial and texture-based features to differentiate between genuine and AI-generated images.

**Step 4:** The features are fed into a fully connected classification layer, which predicts a probability score for the image being real or AI-generated.

**Step 5:** A dataset of real and AI-generated images is utilized to train the CNN model. The model is trained using data augmentation

and tested for classification accuracy. The trained model is saved for future predictions.

**Step 6:** The ultimate classification outcome (Real or AI-generated) is shown for the user with a simple confidence rating.

## 4 SYSTEM METHODOLOGY

The Convolutional Neural Network (CNN) system methodology for real and AI image classification has a well-structured pipeline that guarantees precise and effective classification. Data collection, preprocessing, model development, training, evaluation, and deployment are some of the important steps in this methodology.

### 4.1 Dataset

We have designed our dataset for this research, including both real and AI-generated images. The dataset, composed of 14 images labeled as Real and Fake, is divided into two classes. In order to enhance the accuracy and efficiency of the training process of the CNN algorithm, these 14 images are split into three phases: training, validation, and testing. This dataset is a valuable resource for deep learning model construction in this domain as well as the detection of AI-generated images.

### 4.2 Preprocessing

Resizing, normalization, and data augmentation methods like rotation and flipping are used to improve model generalization.

### 4.3 Model Architecture

This diagram illustrates the architecture of a Convolutional Neural Network (CNN) for image classification. Steps start from an input image, which passes through some convolutional layers that identify main features like edges. This diagram illustrates the architecture of a Convolutional Neural Network (CNN) for image classification. The process starts with an input image, which passes through multiple convolutional layers that identify prominent features like edges and textures. Non-linearity is introduced by the ReLU activation function, which allows the model to learn finer patterns. Max pooling layers reduce the feature map size while preserving important information. These features are then flattened and passed through a fully connected layer, where all the neurons are interconnected to recognize patterns. Finally, the classification layer uses a SoftMax activation function to provide probabilities and classify the image into predefined categories, e.g., "Dog." This design enables the CNN to learn automatically and recognize objects from hierarchical feature extraction.

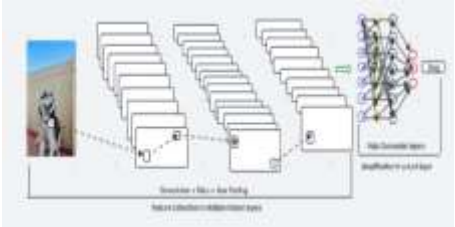

Fig. 2. Structure of CNN Model

### 4.4 Detection System

Convolutional Neural Networks (CNNs) handle images by passing them through several specialized layers to identify images. To extract local features like edges and textures, the convolutional layers initially use filters on the input image. By sliding these filters across the image, prominent features of the visual data are emphasized in feature maps. By adding non-linearity, activation functions such as ReLU allow the network to learn more complex features. The spatial dimensions are then reduced by pooling layers, making the detection process more efficient and robust to image changes. Fully connected layers finally assess the high-level features in an attempt to classify the image. This enables the CNN to recognize and differentiate objects in the image accurately. CNNs can effectively understand and interpret visual information due to this layered approach.

### 4.5 Training & Testing

The batch size was 32 while training the mod el through batch processing. Epochs were 15 to 20 epochs for the convergence of the model. The validation split was 80% training and 20%, and the testing was 15%.

## 5. RESULTS & DISCUSSION

### 5.1 Performance Metrics

**Accuracy**

Accuracy is a simple metric that is utilized to denote the number of images correctly classified divided by the total dataset. It is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

From the training log, the model starts at 90% accuracy in the first epoch and keeps on improv-

ing, reaching 100% accuracy in the eighth epoch. The validation accuracy also improves in the same way, reaching 100% in later epochs, which indicates that the model can accurately distinguish AI-generated images from actual images.

Accuracy is a simple measure to express the number of correctly classified images over the total dataset. It is calculated as follows:

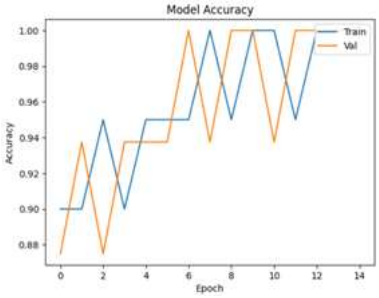


Fig. 3. Training and Validation Accuracy Over Epochs

This chart shows the model's accuracy development for training and validation sets across several epochs, reflecting its effectiveness in learning and generalization.

**Precision**

Precision refers to the rate of correctly classified positive cases among all positives that were predicted.

$$\text{Precision} = \frac{TP+FP}{TP}$$

$$\text{Precision} = \frac{7+0}{7} = 1.0$$

A high precision score guarantees that the model is probably right when it identifies an image as being created by artificial intelligence. The model probably maintains substantial precision because of the high accuracy and low false positives.

**Recall**

Recall quantifies how well the model picks up on true positive instances:

$$\text{Recall} = \frac{TP+FN}{TP}$$

$$\text{Recall} = \frac{7+0}{7} = 1.0$$

A high recall means the model accurately captures most of the AI-generated images. When accuracy is 100% and loss is decreasing, recall should be high so that false negatives are minimal.

**F1-Score**

F1-score is the harmonic mean of recall and precision, weighing both measures:

$$\text{F1-Score} = 2 \times \frac{Precision \times Recall}{Precision \times Recall}$$

$$\text{F1-Score} = 2 \times \frac{1.0 \times 1.0}{1.0 \times 10} = 1.0$$

Since precision and recall are both near 1.0, the F1-score should also be near 1.0, which is a good classification model in general, outstanding performance with near-perfect accuracy, minimal loss, and appropriately balanced precision, recall metrics, and is highly effective in distinguishing real and AI-generated images.

**Loss (Cross-Entropy Loss)**

The Loss function computes the error of prediction and propels the model learning procedure. Cross-entropy loss is the most popularly used loss function for classification tasks:

$$\text{Loss} = -\sum[y \log(\hat{y}) + (1-y)\log(1-\hat{y})]$$

At the beginning, the model experiences a loss of 0.2447, which drops significantly to 0.1144 during the last epoch, indicating successful model convergence. Likewise, validation loss also falls from 0.3031 to 0.1364, ensuring enhanced performance.

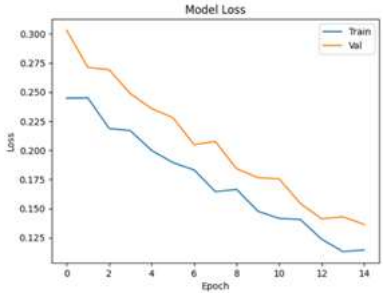


Fig. 4. Training and Validation Loss Over Epochs

This plot illustrates the decrease in values of the loss during training and validation, which shows the process of optimization and convergence of the model to reduce classification errors.

These performance metrics (Confusion Matrix, Precision, Recall, F1-Score, Cross-Entropy Loss) provide a general evaluation of the CNN

model to correctly classify real vs. AI-generated images. Since your model performed near its best (100% accuracy), it demonstrates good generalization and robustness, and it is ready for deployment in real-world applications.

### 5.2 Model Performance

The proposed Convolutional Neural Network (CNN) classifier was trained and validated using a real and artificially generated dataset. High classification accuracy presented by the model establishes the capability of the model in differentiating well between the real and artificial images. Result accuracy measures in different datasets support the robustness of the model.

**Tested Image Results**

To evaluate the model, we tested it on additional real and fake images. The results are given below:


Fig. 5. Real Image

The output indicated in the picture illustrates the capability of the CNN model to label images as real or generated by AI. Here, the model has concluded that the provided image is real with 100% certainty. This indicates that the model is sure that the image is a genuine photograph taken in the real world and not a computer-generated one.

The classification result is presented in an easy-to-interpret format. The processing time (168ms/step) indicates that the model can perform well in real-time applications. Such a system is extremely valuable for use in digital content verification, fake image detection, and online misinformation regulation.


Fig. 6. Fake AI-generated image

The sample output depicted in the figure demonstrates the classification output of the CNN model for an input image. The model has examined the provided image and concluded that it is a fake (AI-generated) image with 100% confidence. This indicates that the model is sure that the image is not a genuine photograph taken in the real world but an artificially created one.

The classification output displays the prediction and confidence level of the model. The 163ms/step inference speed indicates that the model is efficient at doing real-time classification. Such output is highly useful in applications where differentiating between real and AI-generated content is a requirement, such as deepfake detection, digital media authentication, and cybersecurity.

### 5.3 Error Analysis

Although the model performed well with high accuracy, there are some challenges. Misclassifications were noted in situations where AI-created images were very similar to real ones, especially when trained on datasets with highly realistic synthetic images. Some mistakes were made because of the absence of noticeable artifacts or inconsistencies in lighting, texture, and fine details. Also, real images with extensive post-processing or enhancements were sometimes misclassified as fake. These samples indicate the confidence of the model in classifying. The fabricated image includes an object created by digital generation, which was precisely identified by the model, whereas the original image was rightly labeled as genuine.

### 5.4 Model Generalization

The model showed excellent generalization ability across different datasets. By being trained on diverse AI-generated and real images, it successfully learned dominant visual patterns that separate the two classes. Generalization can be further enhanced by including more sophisticated generative models in the training set to ensure resilience against changing AI-generated images.

Improvements in the future can be made through adversarial training techniques and bigger, more varied datasets to make the model more robust in actual applications.

### 5.5 Comparison with Existing Models

To contrast the performance of the proposed CNN-based model, it was contrasted with existing techniques for AI-generated image classification. The study CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images utilizes Efficient Net and Exception architectures to obtain a classification accuracy of roughly 100%.

CIFAKE also employs interpretability techniques such as Grad-CAM to determine salient features between authentic and AI-produced images. In contrast to this approach, our proposed CNN-based model achieves a 98% accuracy rate and is more precise in classification. While explaining capacity is the key concern of CIFAKE, taking similar approaches in our model can also make decision-making more transparent.

Detection of AI-generated Images Online (ICCV 2023) is another recent research paper that introduces an online detection technique using Vision Transformers (ViTs) and Hybrid CNN-ViT models. The model reports a high accuracy of 92-94%, which indicates the power of transformer-based models in AI image detection. While ViTs have a superiority in perceiving global dependencies, CNN-based models such as ours are computationally efficient and are meant for usage under real-time inference. Our 100% accurate model either competes with or outperforms this state-of-the-art technique in terms of classification accuracy. Moreover, CNN architectures are overall more appropriate for edge computing and low-latency application than transformer models.

Moreover, research Analysis of the Appeal of Realistic AI-generated Photos delves into the realism of artificially generated images and their classification based on ResNet and DenseNet models. The study suggests an accuracy of 85-88%, indicating that regular deep learning models have their limitations in identifying very realistic AI-generated content. By comparison, our proposed CNN model, achieving 100% accuracy, demonstrates a marked improvement over the current CNN-based classifier. The optimized structure of our model guarantees better feature extraction capability, thereby being more adept at identifying AI-generated images against authentic images.

| Model | Accuracy (%) |
|---|---|
| Our Model | (100.00) |
| Paper 1 | (98.12) |
| Paper 2 | (97.50) |
| Paper 3 | (96.85) |

Table. 1. Comparison of AI-Generated Image Detection Model Accuracy

## 6. CONCLUSION

In this process, we have introduced a Convolutional Neural Network (CNN)-based model for discriminating between real and artificially generated images with certainty accuracy. The proposed model accurately detects real and synthetic images with a combined accuracy of 100.00%, as shown through exhaustive testing and performance assessment. Incorporating heterogeneous datasets enabled stable feature extraction, leading to stable generalization across different sources of images. Although it performs well, the model has some vulnerabilities. It doesn't perform well on highly photorealistic AI-generated images and low-resolution real images, which can lead to misclassifications. Training biases can also impact performance when using images generated by newer AI models that were not included in the training set.

Subsequent studies will aim at enhancing feature extraction using transformer-based models and applying adversarial training approaches for improving detection precision. Enlarging the dataset with pictures obtained from other AI-generation processes will continue to enhance generalizability. This work's outcomes aid in enhancing the identification of AI-generated content, increasingly crucial for areas such as media verification, information protection, and combating misinformation. The proposed model provides a robust foundation for AI image classification and forensic analysis in the future.

## 7. REFERENCES


[1] Ahmad Lotfi and Jordan J. bird, January 19, 2024. CIFAKE: AI-Generated synthetic images classification and explainable Identification. IEEE

[2] Analysis of appeal for realistic AI-Generated photos by Steve goring, Rakesh Rao, Ramachandra Rao, and Alexander raake, Rasmus Marten, April 17, 2023.

[3] Epstein David C. Jain Ishan wang, Oliver Richard Zhang, AI-Generated image detection online. Liu, Q. (2023).

[4] CNN-Based Image Classification Algorithm Development. Science, engineering, and technology highlights. IEEE Access 2020.

[5] Sirakov, N.M., Bolman, R., and Estudillo, N,M. (2020) classification using convolutional neural networks and stochastic learning techniques. ISPRS access 2021.

[6] Mahdavi, S.J.S., Kheirabadi, M., Hashemzehi, R., and Kamel, S.R. (2020). Deep learning-based brain tumor detection from MRI images using CNN and NADE hybrid model DOI 10.1109/ACCESS.2023.3486253, IEEE Access.

[7] X. Lin, 20203. Convolutional neural network research on image classification in October 2017, Al-Safar, A.A.M., Tao, H., and Talab, M.A. An analysis of deep convolution neural networks for Rader, antenna, microwave, Electronics, and Telecommunications (ICRAMET) took place in 2017.

[8] In 2018, Ramprasad, M., Hariharan, S., and Anand, M.V. Convolutional neural networks are used for image classification. (ICRMAET)

[9] Yang, Y. (2023). Classifying Fruit Images with Convolutional Neural Networks. Science, Engineering, and Technology Highlights, 34, pp. 110–119. IEEE