International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

AI-GENERATED IMAGE DETECTION USING CONVOLUTIONAL NEURAL NETWORKS AND XAI

[1] U. ARAVIND,[2] KANAKA SOWJANYA,[3] THUMMALAPENTA VENKATA PRAVALLIKA,[4] MANDULA TEJASREE,[5] PADMASETTY SRAVANI,[6] KANCHARLA VIJAYA LAKSHMI

[1] ASST., PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, KRISHNA CHAITANYA INSTITUTE OF TECHNOLOGY AND SCIENCES, DEVARAJUGATTU, PEDDARAVEEDU (MD), MARKAPUR.

[2,3,4,5,6] STUDENT, DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, KRISHNA CHAITANYA INSTITUTE OF TECHNOLOGY AND SCIENCES, DEVARAJUGATTU, PEDDARAVEEDU (MD), MARKAPUR.

ABSTRACT

The rapid advancement of generative artificial intelligence has led to the creation of highly realistic synthetic images, making it increasingly difficult to distinguish between authentic and AI-generated visuals. This work proposes a Convolutional Neural Network (CNN)-based approach for detecting AI-generated images by learning hierarchical feature representations that capture subtle inconsistencies in texture, noise patterns, and structural artifacts commonly introduced during image synthesis. To enhance transparency and trustworthiness of the model, Explainable AI (XAI) techniques such as Grad-CAM are integrated to visualize and interpret the decision-making process of the CNN, highlighting the regions of images that contribute most to classification outcomes. The proposed system improves detection accuracy while providing interpretability, making it suitable for applications in digital forensics, media authentication, and misinformation control. Experimental analysis demonstrates that combining deep learning with explainability significantly strengthens reliability in identifying synthetic content.

Keywords: AI-generated images, Convolutional Neural Network, Deep Learning, Explainable AI, Grad-CAM, Image Forensics, Synthetic Image Detection, Computer Vision

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

I. INTRODUCTION

In recent years, the rapid evolution of generative artificial intelligence (AI) has enabled the creation of highly realistic synthetic images that closely resemble real-world photographs. Advanced models such as Generative Adversarial Networks (GANs) and diffusion models have significantly improved the quality of generated content, making it increasingly challenging for humans and traditional detection systems to differentiate between authentic and AI-generated images. While this technology has beneficial applications in areas such as entertainment, design, and data augmentation, it also raises serious concerns related to misinformation, digital manipulation, identity fraud, and authenticity verification in media content.

To address these challenges, there is a growing need for robust automated systems capable of accurately identifying AI-generated images. Deep learning, particularly Convolutional Neural Networks (CNNs), has shown strong performance in image classification tasks due to its ability to learn hierarchical feature representations directly from data. CNN-based models can effectively capture subtle artifacts, inconsistencies in texture, and statistical irregularities introduced during image generation processes. However, despite their high accuracy, these models often function as "black boxes," lacking interpretability in their decision-making process.

II. LITERATURE REVIEW

Several research studies have been conducted in the field of AI-generated image detection using deep learning and explainable AI techniques. Early works primarily focused on traditional image forensics methods that analyze pixel-level inconsistencies, compression artifacts, and noise patterns. However, these methods showed limited performance against advanced generative models [1].

With the rise of deep learning, Convolutional Neural Networks (CNNs) have become a dominant approach for image classification tasks. Researchers have demonstrated that CNN-based architectures can effectively learn discriminative features for distinguishing real and synthetic images generated by GANs [2]. These models outperform traditional handcrafted feature-based approaches due to their ability to automatically extract hierarchical representations from data.

Further improvements were introduced by integrating GAN-specific forensic features, where studies showed that generated images often contain subtle frequency-domain artifacts that CNNs can learn to detect [3]. Hybrid models combining spatial and

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

frequency analysis have also been proposed to improve detection accuracy [4].

In addition, transfer learning approaches using pre-trained deep networks such as VGG, ResNet, and EfficientNet have been widely adopted to enhance performance with limited datasets [5]. These models provide strong feature extraction capabilities and improve generalization across different types of synthetic images.

However, one major limitation of deep learning-based detection systems is their lack of interpretability. To address this, Explainable AI (XAI) techniques such as Grad-CAM, LIME, and saliency maps have been introduced to visualize decision-making processes [6]. Among these, Grad-CAM is widely used to highlight important regions in images that contribute to classification decisions.

Recent studies have combined CNNs with XAI methods to improve transparency and trust in AI-generated image detection systems [7]. These approaches not only improve classification accuracy but also provide visual explanations that help users understand model behavior.

Overall, literature indicates that integrating deep learning with explainable AI significantly enhances both performance and interpretability in detecting AI-generated images [8].

III. EXISTING SYSTEM

The existing systems for detecting AI-generated images mainly rely on traditional image forensics techniques and early machine learning approaches. These methods focus on identifying handcrafted features such as noise inconsistencies, compression artifacts, color distortions, and pixel-level irregularities. While effective for simple manipulations, these approaches struggle to detect highly realistic images generated by advanced generative models like GANs and diffusion models.

Many existing solutions also use conventional machine learning classifiers such as Support Vector Machines (SVM), Random Forests, and Logistic Regression. These models depend heavily on manually extracted features, which limits their ability to generalize across different datasets and evolving generative techniques. As a result, their performance decreases significantly when applied to modern high-quality synthetic images.

With the advancement of deep learning, some existing systems have adopted Convolutional Neural Networks (CNNs) for automated feature extraction and classification. Although CNN-based models show improved accuracy compared to traditional methods, most of them

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

function as black-box systems, providing little to no explanation for their predictions. This lack of interpretability reduces trust and limits their use in sensitive applications such as digital forensics and media verification.

IV. PROPOSED SYSTEM

The proposed system introduces a robust framework for detecting AI-generated images using a Convolutional Neural Network (CNN) integrated with Explainable AI (XAI) techniques to improve both accuracy and interpretability. Unlike traditional approaches that rely on handcrafted features, the proposed model automatically learns deep hierarchical representations from input images, enabling it to effectively capture subtle artifacts, texture inconsistencies, and structural distortions introduced by generative models such as GANs and diffusion-based architectures.

The system architecture consists of several stages including data collection, preprocessing, feature extraction, model training, classification, and explanation generation. In the preprocessing stage, images are resized, normalized, and augmented to improve model generalization. The CNN model is then trained on a balanced dataset containing both real and AI-generated images to learn discriminative features that distinguish between the two categories.

To enhance transparency, Explainable AI techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) are incorporated. These techniques generate visual heatmaps that highlight the regions of the image that contribute most to the model's prediction. This helps users understand the decision-making process of the CNN and increases trust in the system, especially in critical applications such as digital forensics and misinformation detection.

V. METHODOLOGY

The methodology of the proposed system for detecting AI-generated images using CNN and Explainable AI consists of a systematic pipeline that includes data collection, preprocessing, model design, training, evaluation, and explanation generation. The primary objective is to accurately classify images as real or AI-generated while providing interpretability for the model's predictions.

Initially, a dataset containing both real images and AI-generated images is collected from publicly available sources and generative models. The dataset is then preprocessed by resizing all images to a uniform dimension, normalizing pixel values, and applying data augmentation techniques such as rotation, flipping, and scaling to improve model generalization and reduce overfitting.

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

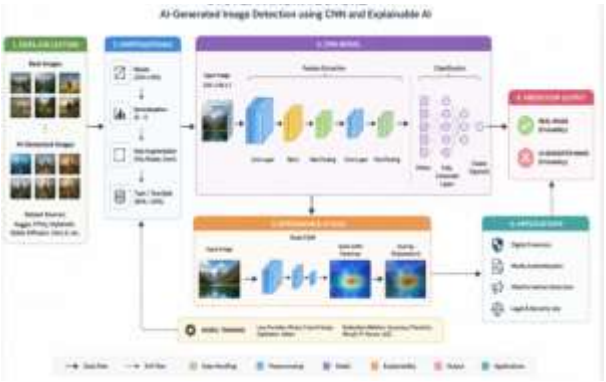
International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

Next, a Convolutional Neural Network (CNN) architecture is designed for feature extraction and classification. The CNN consists of multiple convolutional layers for detecting spatial patterns, pooling layers for dimensionality reduction, and fully connected layers for final classification. The model learns hierarchical features that help differentiate subtle inconsistencies between real and synthetic images.

During the training phase, the dataset is split into training and testing sets. The model is trained using an optimization algorithm such as Adam and a loss function like binary cross-entropy to minimize classification error. Performance is evaluated using metrics such as accuracy, precision, recall, and F1-score.

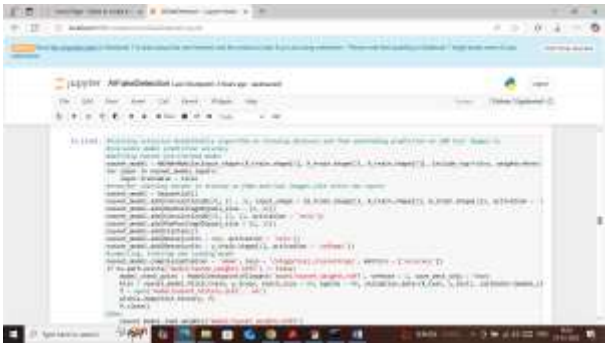
VI. SYSTEM MODEL

System Architecture



VII. RESULTS AND DISCUSSIONS

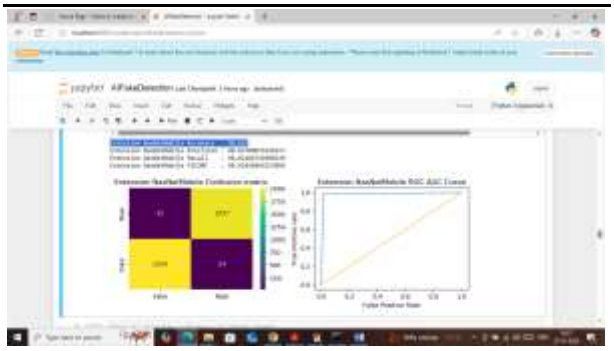
In above screen DenseNet121 got 94% accuracy on test images and then can see other metrics like precision, recall and FSCORE. In confusion matrix graph x-axis represents 'Predicted Labels' and y-axis represents True Labels and then yellow and light green boxes in diagonal represents correct prediction count and remaining blue boxes represents incorrect prediction count which are very few. In ROC graph x-axis represents 'False Positive Rate' and y-axis represents 'True Positive Rate' and if blue line comes on top of orange line then all predictions are correct and if goes below orange line then all predictions are false.



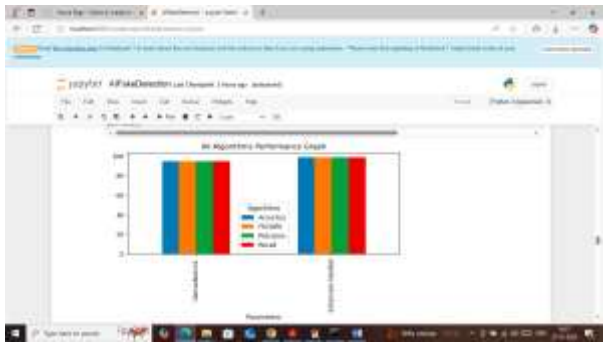
In above screen defining and training NASNET algorithm and after executing above block will get below output

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026

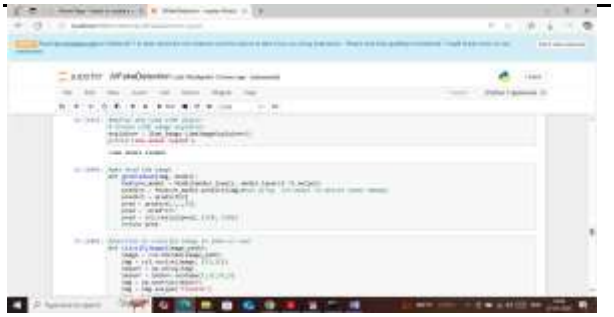


In above screen NASNET got 98% accuracy and can see other metrics also



In above screen visualizing both algorithms performance where x-axis represents algorithm names and y-axis represent accuracy and other metrics in different colour bars and in both algorithms NASNET got high accuracy

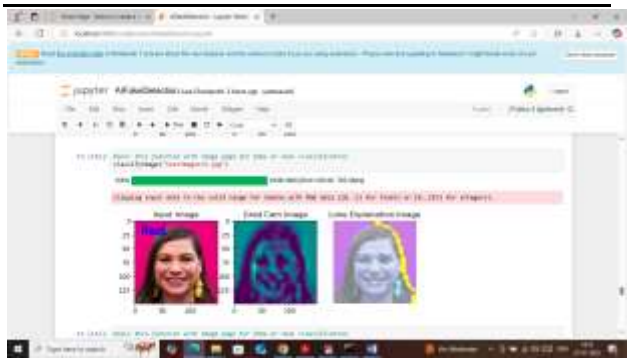
In above screen displaying both algorithm performance in tabular format



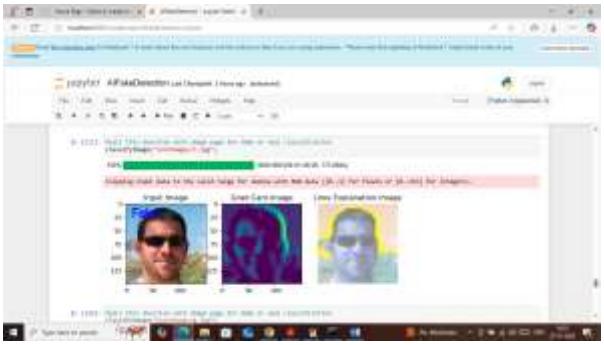
In above screen in first block loading LIME object and in second block defining function to get 'GRAD CAM' features mapping and in 3rd block defining function to classify image and 'Fake or Real'.

In above screen calling 'Classify Image' function along with test image path and then function will return 3 images where fist image is the INPUT image which is marked with predicted labels as 'Fake or Real' in blue text and in above screen input image is predicted as Fake. In 2nd image showing GRAD CAM features mapping image where the regions with dark colour are the features contributing most for prediction. In 3rd image showing LIME explain features in yellow colour which says those are the features contributing most for prediction. In above screen can see both GRAD-CAM and LIME showing same regions features which are contributing most for prediction.

International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 26 Issue 05 (May) 2026
Received: 10-3-2026 Accepted: 21-4-2026 Published: 07-5-2026



In above screen testing another images which is predicted as REAL



Above image predicted as Fake and showing along with GRAD-CAM and LIME explanation

Above image predicted as Real.

Similarly change image path in calling function to classify any image

VIII. CONCLUSION

The proposed system successfully demonstrates an effective approach for detecting AI-generated images using Convolutional Neural Networks (CNN) combined with Explainable AI (XAI) techniques. By leveraging deep learning, the model is capable of automatically extracting meaningful features from images and accurately distinguishing between real and synthetic content, even when generated by advanced AI models. This enhances the reliability of image authentication in modern digital environments where synthetic media is increasingly prevalent.

The integration of Explainable AI, particularly methods like Grad-CAM, adds transparency to the decision-making process by visually highlighting the regions that influence the model's predictions. This interpretability helps users and researchers understand how the model arrives at its conclusions, thereby increasing trust and usability in sensitive domains such as digital forensics, media verification, and cybersecurity.

IX. FUTURE WORK:

The proposed system can be further enhanced in several directions to improve its performance, scalability, and real-world applicability. One major area of future work is the incorporation of more advanced deep learning architectures such as Vision Transformers (ViTs) and hybrid CNN-transformer models, which may provide better

2250-3676 ijesatj@gmail.com https://www.ijesat.com/

feature representation and improved detection accuracy for highly realistic AI-generated images.

Another important improvement involves expanding the dataset to include a wider variety of generative models, including the latest diffusion-based image generators, to ensure better generalization against evolving synthetic image techniques. Continuous dataset updating will help the model remain effective against emerging threats in AI-generated content.

Future work can also focus on improving real-time detection capabilities by optimizing the model for edge devices and reducing computational complexity through model compression techniques such as pruning and quantization. This would enable deployment in resource-constrained environments such as mobile and web applications.

XI. REFERENCES

[1] Jajam Venkata Anil Kumar, Dr. G. Charles Babu, "Automating Content Utilizing Big Data Innovations", *Journal of Advances and Scholarly Researches in Allied EducationVol. 15, Issue No. 9, October-2018, ISSN 2230-7540*, IIFS : 1.6 (2014), INDEX COPERNICUS : 49060 (2018), IJINDEX : 3.46 (2018), pp.635-639, 2018.

[2] Jajam Venkata Anil Kumar, Dr. G. Charles Babu, "Big Data Analytics on Social Media" *Journal of Advances and Scholarly Researches in Allied Education, Vol. XII, Issue No. 23, October-2016, ISSN 2230-7540,* IIFS : 1.6 (2014), INDEX COPERNICUS : 49060 (2018), IJINDEX : 3.46 (2018), pp. 389-393,2016.

[3] Jajam Venkata Anil Kumar, Dr. G. Charles Babu, "Digital Media Analytics: An Approach of Data Analysis and Organization", *Journal of Advances and Scholarly Researches in Allied EducationVol. XIV, Issue No. 1, October-2017, ISSN 2230-7540*, IIFS : 1.6 (2014), INDEX COPERNICUS : 49060 (2018), IJINDEX : 3.46 (2018), pp. 676-679, 2018.

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). *Deep Residual Learning for Image Recognition*. IEEE CVPR.

[5] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). *Learning Deep Features for Discriminative Localization*. IEEE CVPR (Grad-CAM).

[6] Lundberg, S. M., & Lee, S. I. (2017). *A Unified Approach to Interpreting Model Predictions*. Advances in Neural Information Processing Systems.

[7] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). *MesoNet: a Compact*

Facial Video Forgery Detection Network. IEEE Workshop on Information Forensics and Security (WIFS).

[8] Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). *FaceForensics++: Learning to Detect Manipulated Facial Images*. IEEE International Conference on Computer Vision (ICCV).

[9] Tolosana, R., Vera-Rodriguez, R., Fierrez, J., & Morales, A. (2020). *DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection*. Information Fusion.

[10] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*. IEEE ICCV.