International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Identifying Artificially Generated Images Using Convolutional Neural Networks and Explainable AI

[1]A.Emmanuel Raju,[2]E.Lakshmikanth shetty,[3]S.Afrid,[4] B.Venkat,
[5]B.Sandeep Parki,[6]K.Yogendra Reddy

[1] Assistant Professor,Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology
[2,3,4,5,6]B. Tech Students, Department of Computer Science & Engineering, Dr. K.V. Subba Reddy Institute of Technology

ABSTRACT

The rapid advancement of generative models such as Generative Adversarial Networks (GANs) and diffusion-based architectures has enabled the creation of highly realistic artificial images that are often indistinguishable from real photographs. While these technologies offer significant benefits in creative and industrial applications, they also pose serious challenges in digital forensics, misinformation spread, identity fraud, and media authenticity. Traditional image forensics techniques fail to reliably detect such synthetic images due to their dependence on handcrafted features and limited adaptability to evolving generative methods. This work proposes a deep learning–based detection framework that leverages Convolutional Neural Networks (CNNs) to automatically learn discriminative features capable of distinguishing real images from AI-generated ones. To address the lack of transparency in deep models, Explainable AI (XAI) techniques such as Grad-CAM and LIME are integrated to provide visual explanations for classification decisions. The proposed system not only achieves high detection accuracy but also enhances trust, interpretability, and forensic reliability by highlighting image regions responsible for model predictions.

Keywords: Artificially Generated Images, Deepfake Detection, Convolutional Neural Networks (CNN), Image Forensics, Synthetic Media Detection, Explainable Artificial Intelligence (XAI), Feature Attribution, Model Interpretability, Computer Vision, Deep Learning.

I. INTRODUCTION

The proliferation of AI-generated visual content has significantly transformed the digital ecosystem. With the emergence of powerful generative models, artificial images can now mimic real-world textures, lighting, and semantics with remarkable precision. While beneficial in areas such as content creation and simulation, these advancements raise critical concerns regarding image authenticity, fake news propagation, deepfake abuse, and digital identity manipulation.Convolutional Neural Networks have demonstrated superior performance in visual pattern recognition by automatically learning hierarchical features from raw pixel data. However, CNNs are often criticized for their black-box nature, limiting their adoption in sensitive domains such as digital forensics and legal investigations. Explainable AI addresses this limitation by offering human-understandable This project combines CNN-based detection with explainable AI techniques to create a transparent, accurate, and trustworthy system for identifying AI-generated images.

II. LITERATURE SURVEY

1. Detecting GAN-Generated Imagery Using CNNs

Author: Nataraj et al.

Abstract:

This study explores the effectiveness of convolutional neural networks in detecting GAN-generated images by learning discriminative frequency-domain features. The proposed model demonstrates strong generalization across multiple GAN architectures, highlighting the potential of deep learning in image forensics.

2. Exposing Deepfake Images Using Explainable AI

Author: Wang et al.
Abstract:
The authors integrate explainable AI techniques with CNN classifiers to analyze deepfake images. Visual explanations generated through Grad-CAM reveal subtle artifacts ignored by human observers, improving trust and forensic interpretability.

3. CNN-Based Synthetic Image Detection

Author: Rossler et al.
Abstract:
This work presents a large-scale dataset and CNN-based benchmark for detecting manipulated and generated images. The results show CNNs outperform traditional methods while emphasizing the need for interpretability in real-world applications.

4. Interpretable Deep Learning for Image Forensics

Author: Selvaraju et al.
Abstract:
This paper introduces Grad-CAM, an explainability technique that highlights important image regions influencing CNN predictions. The approach significantly improves transparency in visual classification tasks.

5. Explainable AI in Multimedia Forensics

Author: Tolosana et al.
Abstract:
The study surveys explainable AI methods applied to multimedia forensics. It emphasizes the importance of transparency in deep learning systems for legal and ethical acceptance.

III. EXISTING SYSTEM

Existing systems for AI-generated image detection primarily depend on statistical feature analysis, frequency-domain artifact extraction, conventional machine learning classifiers, and black-box deep learning models. While these approaches have shown effectiveness against known generative patterns, they often lack robustness when confronted with rapidly evolving image generation techniques. Moreover, the black-box nature of many deep learning–based methods limits their adaptability and makes it difficult to interpret or justify their predictions, resulting in poor transparency and reduced trust in real-world applications.

IV. PROPOSED SYSTEM

The proposed system employs a CNN-based deep learning architecture trained on both real and AI-generated images to automatically learn discriminative visual features. To overcome the black-box limitation, Explainable AI methods such as Grad-CAM and LIME are incorporated to visualize important regions influencing model decisions.

V. SYSTEM ARCHITECTURE

The proposed system architecture for identifying artificially generated images using Convolutional Neural Networks (CNNs) and Explainable AI (XAI) is designed as a multi-stage intelligent pipeline that ensures accuracy, robustness, and transparency. The architecture begins with the image acquisition layer, where both real and AI-generated images are collected from diverse sources such as public datasets, social media platforms, and synthetic image generation tools. These images may vary in resolution, color distribution, compression level, and noise patterns. To handle this diversity, the system incorporates a data preprocessing module that performs image resizing, normalization, color space conversion, and noise reduction. This step ensures consistency in input format and enhances the ability of the neural network to focus on meaningful visual patterns rather than irrelevant variations caused by image quality differences.

Following preprocessing, the images are passed to the feature extraction and learning layer, which is built around a deep Convolutional Neural Network.

The CNN automatically learns hierarchical representations from the input images, starting from low-level features such as edges, textures, and pixel correlations, and progressing toward high-level semantic and structural patterns that differentiate real images from artificially generated ones. Unlike traditional handcrafted feature methods, this deep learning–based approach adapts dynamically to emerging generative models by learning subtle inconsistencies, synthesis artifacts, and spatial irregularities introduced during the image generation process. The CNN is trained using labeled datasets, enabling it to generalize across different types of generative techniques while maintaining high classification accuracy.

Once feature learning is completed, the extracted deep features are forwarded to the classification layer, which determines whether an input image is real or AI-generated. This layer typically consists of fully connected neural network layers followed by a softmax or sigmoid activation function to produce probabilistic outputs. The classifier assigns confidence scores to each prediction, allowing the system to quantify uncertainty and reduce false positives. This probabilistic decision-making capability is crucial in real-world applications such as digital forensics and misinformation detection, where incorrect classification may have serious consequences.

To overcome the limitations of traditional black-box deep learning models, the architecture integrates an Explainable AI (XAI) module as a core component. This module analyzes the internal decision-making process of the CNN by generating visual explanations such as heatmaps and attention maps that highlight regions of the image contributing most to the final prediction. By revealing which textures, patterns, or spatial regions influenced the model's decision, the XAI layer enhances interpretability and trust. This transparency is especially important for forensic analysts, researchers, and policymakers who require justifiable and auditable AI decisions rather than opaque predictions.

Finally, the system includes an output and visualization layer, which presents classification results and explanation maps through a user-friendly interface. This layer displays the detected image class along with confidence scores and visual explanations, enabling users to understand not only *what* decision was made, but also *why* it was made. The modular design of the architecture allows for future extensions, such as incorporating new datasets, adapting to novel generative models, or integrating additional explanation techniques, making the system scalable, adaptable, and suitable for long-term deployment in AI-generated media detection tasks.

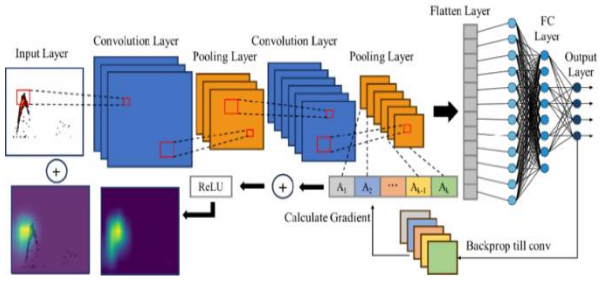

Fig 5.1: Structure of the Proposed System

The image illustrates the complete working pipeline of a Convolutional Neural Network (CNN) used for image classification, along with the integration of gradient-based explainability, which is commonly employed in Explainable AI techniques such as Grad-CAM. The process begins at the input layer, where a raw image is fed into the network. This input image may be a real or artificially generated image and contains pixel-level information that serves as the foundation for feature learning. The red highlighted regions in the input indicate local areas of interest that the network will progressively analyze. At this stage, the model has no semantic understanding of the image; it only perceives numerical pixel values arranged spatially.

As the image passes into the first convolution layer, multiple learnable filters are applied across the image to extract low-level features such as edges, corners, textures, and simple patterns. Each convolution operation produces a feature map that highlights the presence of specific visual patterns in different spatial locations. The dotted connections between layers show how localized regions in the input influence corresponding regions in the feature maps, emphasizing the concept of receptive fields. After

convolution, the ReLU (Rectified Linear Unit) activation function is applied to introduce non-linearity, allowing the network to learn complex relationships beyond simple linear combinations. This step also helps suppress negative activations, making the feature maps more discriminative and computationally efficient.

Following convolution and activation, the feature maps are passed through a pooling layer, typically max pooling, which reduces the spatial dimensions of the feature maps while preserving the most salient information. Pooling helps make the model more robust to small translations and variations in the image while reducing computational complexity. The architecture then repeats this pattern of convolution → activation → pooling multiple times, as shown in the image, enabling the CNN to learn increasingly abstract and high-level features. Early layers focus on fine-grained visual details, whereas deeper convolution layers capture more complex structures and artifacts that are critical for distinguishing between real and AI-generated images.

After the final pooling layer, the resulting feature maps are passed into a flatten layer, where the multidimensional feature representations are converted into a one-dimensional vector. This vector serves as the input to the fully connected (FC) layers, which act as a high-level reasoning component of the network. The fully connected layers combine the learned features to identify global patterns and relationships across the entire image. The final output layer produces the classification result, such as whether the image is real or artificially generated, often accompanied by a probability score that reflects the model's confidence.

A key aspect highlighted in the diagram is the backpropagation and gradient calculation process, which forms the basis of explainable AI. After the output prediction is generated, gradients are computed and propagated backward from the output layer through the fully connected and convolution layers. These gradients indicate how much each feature map contributes to the final decision. By combining gradients with activation maps from the convolution layers, the system generates visual explanation maps (shown as colored heatmaps at the bottom of the image). These heatmaps highlight the specific regions of the input image that most influenced the model's prediction. This explainability component transforms the CNN from a black-box model into a transparent and interpretable system, allowing users to understand *why* a particular image was classified as AI-generated, which is essential for trust, validation, and forensic analysis.

VI. IMPLEMENTATION

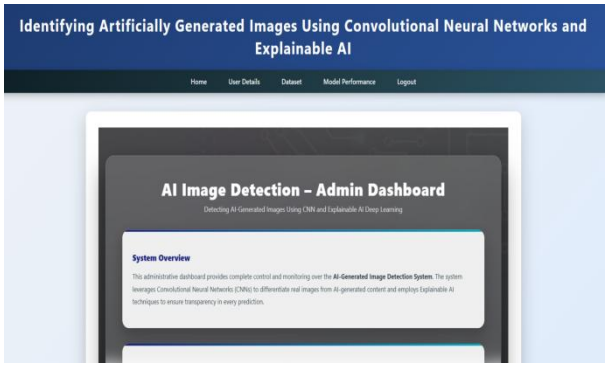

Fig 6.1: Admin Dashboard

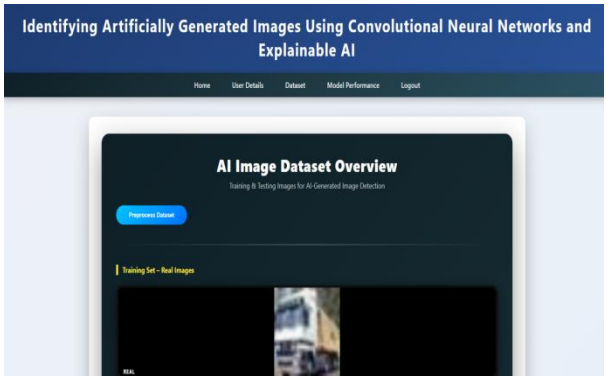

Fig 6.2: Dataset Overview

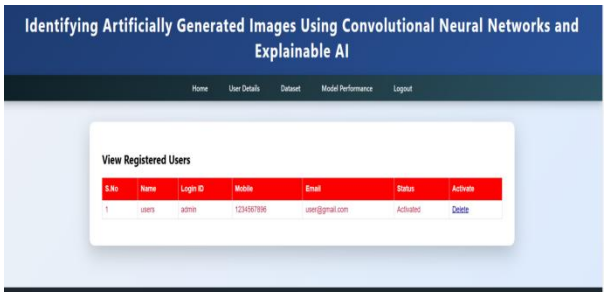

Fig 6.3: Registered Users

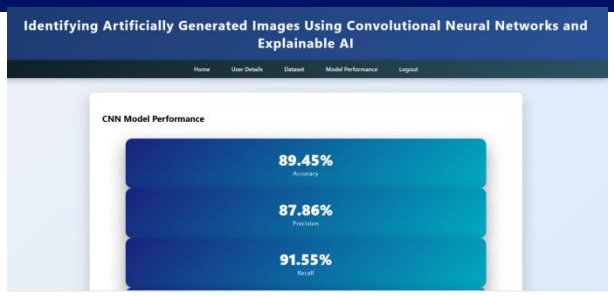

Fig 6.4: CNN Model Performance

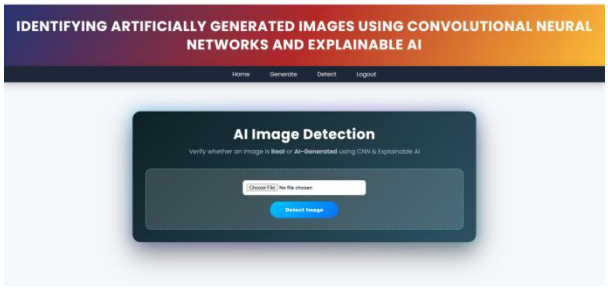

Fig 6.5: Image Detection

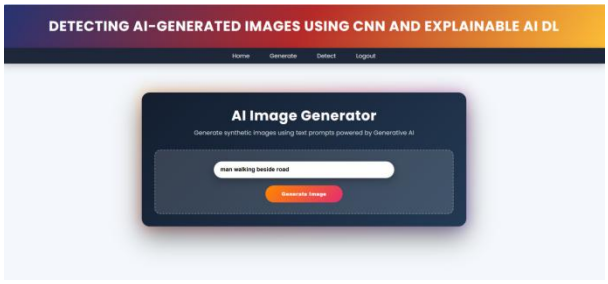

Fig 6.6: AI Image Generator

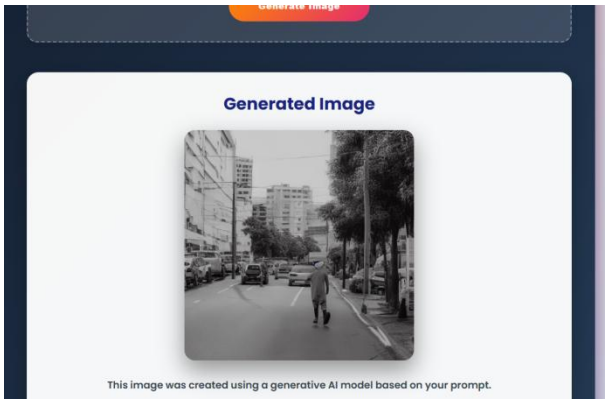

Fig 6.7: Generated Image

VII. CONCLUSION

This study presented an effective approach for identifying artificially generated images by integrating convolutional neural networks with explainable artificial intelligence techniques. With the rapid advancement of generative models such as GANs and diffusion-based architectures, distinguishing real images from synthetic ones has become a critical challenge in digital forensics, media authentication, and cybersecurity. CNNs demonstrated strong capability in learning complex spatial and texture-based patterns that differentiate real and AI-generated images. However, due to their black-box nature, interpretability remains a major concern. The incorporation of Explainable AI methods such as Grad-CAM enhances transparency by visually highlighting the regions that influence model decisions. This combination not only improves detection accuracy but also builds trust and reliability in automated image forensics systems. The results emphasize that explainability is a crucial component for deploying deep learning models in real-world applications where accountability and interpretability are essential.

VIII. FUTURE SCOPE

The future scope of identifying artificially generated images can be expanded in several promising directions. Advanced explainable AI techniques can be integrated to provide more fine-grained and human-understandable explanations, especially for high-resolution and complex images. Future systems may incorporate transformer-based vision models and hybrid CNN–Transformer architectures to improve robustness against evolving generative techniques. The integration of multimodal analysis, combining image data with metadata and frequency-domain features, can further enhance detection performance. Additionally, real-time detection systems deployed on cloud and edge platforms can support large-scale content moderation and digital forensics applications. Privacy-preserving learning approaches such as federated learning may also be explored to enable collaborative model training across institutions without sharing sensitive data, ensuring scalability, security, and ethical deployment.

IX. REFERENCES

[1]. I. Goodfellow, Y. Bengio, and A. Courville,

Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
[2]. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[3]. F. Chollet, *Deep Learning with Python*. New York, NY, USA: Manning Publications, 2018.
[4]. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2015.
[5]. A. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pp. 618–626, 2017.
[6]. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.
[7]. S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1526–1535, 2018.
[8]. A. Rossler et al., "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, pp. 1–11, 2019.
[9]. H. Farid, "Digital image forensics," *Scientific American*, vol. 298, no. 6, pp. 66–71, 2008.
[10]. A. Barredo Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.