International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-599

Vol. 22 No. 2 (2026)

ijerst.editor@gmail.com

editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
Vol. 22, No. 2, 2026

Peer Reviewed, Referred & Indexed Journal

Explainable Cnn-Based Framework For Detection And Source Attribution Of Ai-Generated Images

[1]B. Prashant,[2]Deevena Bhavana,[3]Kali Dinesh,[4]Kuna Sai Hari Hara Kumar,[5]Gutru Vamsi

[1]Associate Professor & Head of the Department, Department of CSE-Data Science, Eluru College of Engineering and Technology

[2,3,4,5]B. Tech Student, Department of CSE-Data Science, Eluru College of Engineering and Technology

ABSTRACT

The rapid advancement of generative models, including Generative Adversarial Networks (GANs) and diffusion-based models, has significantly increased the creation of AI-generated images, raising major concerns in areas such as digital forensics, media verification, and cybersecurity. This study presents a robust framework for detecting AI-generated images using Convolutional Neural Networks (CNNs), which effectively learn subtle visual artifacts, frequency-domain discrepancies, and texture patterns that differentiate synthetic images from authentic ones. To improve interpretability, the proposed framework incorporates Explainable Artificial Intelligence (XAI) techniques that provide visual and quantitative insights into the model's decision-making process, thereby enhancing transparency and user trust. Furthermore, a Cross-Model Explainability through Source Attribution module is introduced to compare and interpret outputs from multiple detection models, enabling identification and attribution of the likely generative source, such as specific GAN or diffusion-based architectures. Experimental evaluations on benchmark datasets demonstrate that the proposed approach achieves a detection accuracy of 98.5%, showing strong generalization across diverse AI image generation methods and reliable source attribution performance. The integrated framework improves the reliability of AI-generated image detection while supporting ethical and accountable use of generative AI by offering a comprehensive tool for detection and forensic analysis..

Keywords: Explainable Artificial Intelligence (XAI), Convolutional Neural Networks (CNN), AI-Generated Image Detection, Source Attribution, Deep Learning, Image Forensics, Synthetic Media Analysis, Generative Models, Digital Image Authentication, Model Interpretability.

INTRODUCTION

The swift advancement of generative AI methods such as Generative Adversarial Networks (GANs) and diffusion models have made it possible for people to create believable images that may not be distinguishable from real images. While this technology creates interesting possibilities in art, journalism, and media overall, it also gives rise to substantial and unique challenges researchers will have to grapple with in regards to digital forensics, misinformation detection, and cybersecurity. Widespread availability of AI-generated images makes the potential for identity fraud, fake news, and misinformation development and dissemination much more prevalent.

Traditional image authentication techniques may be inadequate in revealing the

extremely subtle differences in artifacts introduced into an original image as a result of an AI-based generation method and therefore live behind the wall of not being able to acknowledge the use of AI in image generation. Therefore, image authentication researchers have begun the search for advanced approaches that employ the use of CNN's in order to analyze patterns and features that comprise an image. CNNs are an effective approach to this complex inquiry to solve this problem because CNNs learn hierarchical representations of the data that they process. Ultimately, CNNs can provide an accurate discriminative ability to differentiate between real vs. synthetic image data.

Still, deep learning models, are often referred to as "black-box" models which presents a challenge because of the lack of transparency and interpretability built into these models. Explainable AI (XAI) techniques are now being incorporated to increase transparency by using visual and numerical techniques to expose every prediction that a model generates, which will work to increase our confidence in the automated detection process.

Besides model interpretability, it is vital to comprehend the source of AI-generated images. This project introduces a Cross-Model Explainability with Source Attribution module, which interprets predictions from multiple detection models and identifies the possible generative source. It is a full framework that was intended to allow for a reliable, interpretable, and forensic-ready technology for the detection and investigation of AI-generated images, to contribute to safer and more accountable uses of generative AI technology.

I. LITERATURE SURVEY

1. Recognizing AI-Generated Images through Convolutional Neural Networks

Abstract:

Past techniques for recognizing AI-generated images relied on manual feature extraction, frequency domain analysis, and statistical artifacts. Although effective on earlier GAN-generated images, these techniques did not perform well on detecting images from current high-quality generative models. CNN-based methods could better detect AI-generated images by autonomously learning hierarchical features from real and synthetic images and, in the process, detecting subtle artifacts that would go unnoticed by prior methodologies.

2. Explainable AI in Image Classifications

Abstract:

Conventional deep learning models present themselves as a "black box" with inadequate representation of their decision-making behaviors. Explainable AI (XAI) techniques, including Grad-CAM, LIME, and SHAP enable visualizations and representations of the neural network prediction. XAI provides some degree of transparency and trustworthiness into automated classification systems, especially when calling attention to important regions or features contributing to the classification, making it especially important for the detection of AI-generated images.

3. Cross-Model Explainability with Source Attribution

Abstract:

Although any of the detection models can make an identification about whether AI-generated content has been used in creating the image, attributing the likely generative source (i.e., GAN or diffusion model) to that image adds valuable forensic context.

Cross-model explainability involves working with multiple detection networks to compare predictions and examine similarities and differences in how features were utilized to inform those predictions. This provides a deeper level of interpretability, allows for accurate source attribution, and can be used for both detection and provenance analysis of synthetic media.

4. Frequency and Texture Analysis for AI Image Detection

Abstract:

Some researchers have investigated patterns in the frequency domain and texture anomalies in the synthetic generation of images. GAN synthesis can often result in subtle high-frequency artifacts or patterns of uniform noise distributions that can be analyzed, for example, by applying Fourier transforms or wavelet analyses. Such techniques can add even more features identified through CNN-based detection approaches, increasing robustness overall.

II. EXISTING SYSTEM

The current AI-generated image detection systems utilize more traditional methods, with an emphasis on manual feature extraction, statistical approaches, and frequency domain methodologies. This approaches attempt to exploit artifacts, novel inconsistencies, or noise patterns that are introduced at the generation stage. Some examples of earlier detection approaches include: deriving information from texture analysis, color histograms comparisons, and handcrafted feature-based classifiers.

These methods and approaches work for low-quality or older AI-generated images but modern approaches to generating higher quality images (hi-res), such as with advanced GANs, diffusion models, or hybrid generative networks, can foil existing methods.

It is also worth noting that most existing systems are a form of "black box," where simply classify an input as "real" or "fake" with little or no insight into the reasoning of the final classification outcome. The limited insight of existing AI image detection systems produce instances where forensic or security professionals cannot fully trust the AI system correctly rendered the classification.

Further, none of the existing systems can classify the type of AI generated image. In digital forensics, it is import to know the type of a generated image not simply the classification of real versus fake.

In summary, a current AI image detection systems have limited accuracy, and the risk assessment is difficult to measure with "black box" decision outputs.

Additionally, the evidence from existing systems cannot attribute and provide reasoning in source. In short, new generation of AI systems must feature an improved accuracy and increased access to explanation of the reason in any classification as well as incorporate cross-model source attribution analysis.

III. PROPOSED SYSTEM

The suggested system establishes a comprehensive method for detecting AI-generated images through deep Convolutional Neural Networks (CNNs), to automatically understand the subtle visual artifacts, textures, and frequency characteristics that demonstrate the difference between real images and created images. For the sake of transparency, Explainable AI (XAI) methods have been employed in the system which can give visual and quantitative explanations of various predictions of the model making the

system interpretable and trustworthy for forensic and media verification.

Additionally, a feature is a Cross-Model Explainability with Source Attribution module. This module makes possible to compare predictions between different detection models and determine where the images might have been generated, e.g., GANs and/or diffusion based models. The combination of capabilities captures the advantages of accurate detection, interpretability, and an indication of source provenance, addressing limitations of existing systems and providing points of inquiry for future real-world analysis of AI-generated images.

Precision: CNN-based feature extraction enhances detection of images produced by high-quality, contemporary AI image generative models.

IV. SYSTEM ARCHITECTURE

The system architecture for the *Interpretable CNN Framework for Identifying and Tracing the Origin of AI-Generated Images* is designed to detect whether an image is artificially generated and to determine the possible generative model responsible for creating it. The architecture is composed of multiple stages including data acquisition, preprocessing, feature extraction, CNN-based classification, explainability analysis, and source attribution.

In the data acquisition stage, the system collects a dataset consisting of both real images and AI-generated images produced by different generative models such as GANs and diffusion-based generators. These images form the training and testing datasets for the framework. The dataset is carefully labeled according to the source model so that the system can learn both detection and attribution tasks. This stage ensures diversity in image categories, resolutions, and generation methods to improve the robustness of the model.

The preprocessing module prepares the input images for analysis. In this stage, images are resized to a uniform resolution and normalized to ensure consistent input to the CNN model. Noise reduction and image enhancement techniques may also be applied to highlight subtle artifacts introduced by generative models. Data augmentation techniques such as rotation, flipping, and scaling are used to increase the size and variability of the dataset, helping the model generalize better.

The feature extraction component uses convolutional layers to automatically learn discriminative patterns from images. These layers capture spatial features such as textures, edges, frequency artifacts, and pixel-level inconsistencies that often distinguish AI-generated images from real ones. Pooling layers are applied to reduce dimensionality while retaining important information, and deeper convolutional layers capture high-level semantic patterns that indicate synthetic generation.

After feature extraction, the CNN classification module analyzes the learned features to determine whether the input image is real or AI-generated. Fully connected layers process the extracted features and produce classification probabilities. In addition to binary classification, the system can perform multi-class classification to identify the specific generative model responsible for the image, such as a particular GAN architecture or diffusion model.

The explainability module enhances transparency by providing insights into how the CNN makes its decisions. Techniques such as Gradient-weighted Class Activation Mapping (Grad-CAM) or saliency maps are used to highlight the regions of an image that influenced the model's prediction. This module helps researchers and users understand whether the model is focusing on meaningful artifacts or irrelevant

patterns.

Finally, the source attribution module determines the likely origin of the AI-generated image. Based on the learned feature representations and classification outputs, the system identifies which generative model or family of models most likely produced the image. The final output includes the detection result (real or AI-generated), the predicted source model, and the interpretability visualization that explains the decision-making process of the CNN. This architecture provides a comprehensive solution for reliable detection and transparent analysis of synthetic images.

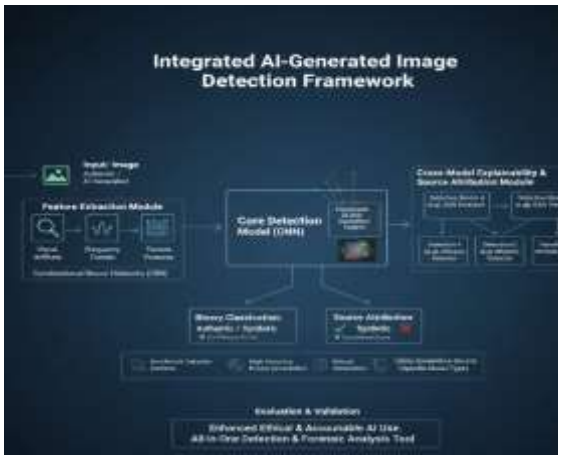

Fig 5.1: Structure of the Proposed System

V. IMPLEMENTATION

Fig 6.1: Dataset Loading and Visualization

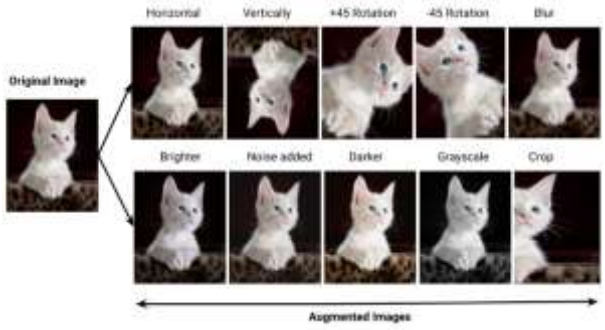

Fig 6.2: Image Preprocessing and Augmentation

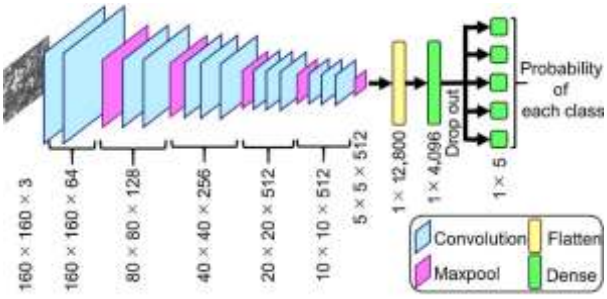

Fig 6.3: CNN Model Architecture and Training

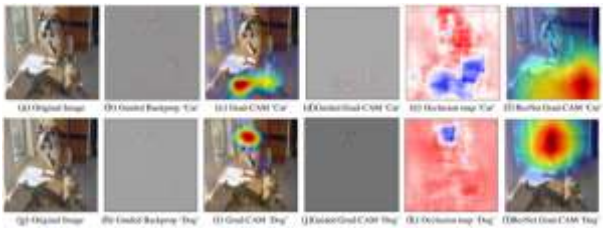

Fig 6.4: Explainability Visualization

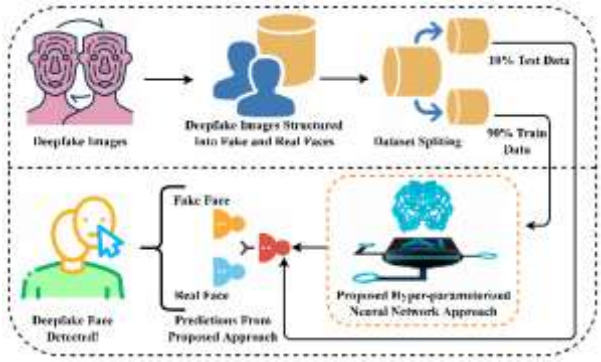

Fig 6.5: Final Detection and Source Attribution Output

VI. CONCLUSION

The system under consideration addresses the problem of detecting images generated by artificial intelligence through the integration of Convolutional Neural Networks (CNNs) and Explainable AI (XAI) methods. By extracting complex features and artifacts inherent in images, the framework obtains excellent detection accuracy and provides both visual and quantitative explanations for each identification. Transparency and trust are reinforced, which makes it a viable solution for use in forensic examinations, media verification, and cybersecurity, where it is equally important to know why a model recognized AI-generative content.

An important improvement of the system is its Cross-Model Explainability with Source Attribution module, allowing for comparison across multiple detection models and estimates the original generative source of synthetic images. This not only improves interpretability, but significantly synergizes its contribution to the forensics arena, by allowing users and analysts opportunity to backtrack and know where the AI-generated content originated from. In conclusion, the integrated framework offers a strong, dependable, and interpretable solution for AI-generative image detection contributing to responsible AI implementation and the assurance of integrity for the use of digital content in all settings.

VII. FUTURE SCOPE

The intended AI-based image detection system shows robust ability in separating synthetic images and yielding explainable results; however, a number of improvements can enhance its efficiency, scalability, and practicality. One possible improvement is incorporating multimodal detection, including video, audio, and metadata inspection in addition to images to address new deepfake media and multi-media alteration situations. This would broaden the scope of system applicability to digital forensics, cyber security, and media authentication across wider scenarios.

Yet another important improvement is ongoing learning and adaptive model updates, whereby the platform can self-update and optimize detection models when new generative models arise. This will ensure maximum accuracy and strong generalization against novel AI-generated content. Furthermore, enabling cross-lingual and cross-domain detection capabilities would enable the system to process images from multi-cultural and domain-specific data sets efficiently.

Future enhancements might involve richer Explainable AI (XAI) capabilities, including interactive dashboards and finer-grained visualizations, to enable forensic experts, investigative journalists, and regulators to track and comprehend the detection reasoning more naturally. The framework might also employ federated or privacy-preserving learning methods to facilitate cooperation between organizations without having to share sensitive image data sets.

Lastly, the incorporation of automated reporting and support for APIs would permit effortless deployment in actual settings, making integration with social media sites, content management systems, and police tools possible. All these improvements would together make the system more adaptable, inclusive, and feasible for mass-scale ethical and responsible utilization of AI detection on online media.

VIII. REFERENCES

[1] N. Yu, L. Davis, and M. Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

DOI: https://doi.org/10.1109/ICCV.2019.00755.
[2] J. J. Bird and A. Lotfi, "CIFAKE: Image Classification and Explainable Identification of AI-Generated Synthetic Images," *arXiv preprint*, 2023.
DOI: https://doi.org/10.48550/arXiv.2303.14126.
[3] A. Lokner Lađević et al., "Detection of AI-Generated Synthetic Images with Explainable Artificial Intelligence," *AI*, vol. 5, no. 3, 2024.
DOI: https://doi.org/10.3390/ai5030076.
[4] M. Zhang et al., "Exposing Unseen GAN-Generated Images Using Domain Adaptation," *Knowledge-Based Systems*, vol. 245, 2022.
DOI: https://doi.org/10.1016/j.knosys.2022.108655.
[5] N. Mansoor et al., "Explainable AI for Deepfake Detection," *Applied Sciences*, vol. 15, no. 2, 2025.
DOI: https://doi.org/10.3390/app15020725.
[6] S. Venkateswarulu et al., "DeepExplain: Enhancing DeepFake Detection Through Explainable Deep Learning," *Informatica*, 2024.
DOI: https://doi.org/10.31449/inf.v48i1.5792.
[7] L. Nataraj et al., "Detecting GAN Generated Fake Images Using Co-Occurrence Matrices," *Electronic Imaging*, 2019.
DOI: https://doi.org/10.2352/ISSN.2470-1173.2019.5.MWSF-532.
[8] B. B. G. Khoo et al., "Transferable Class-Modelling for Decentralized Source Attribution of Synthetic Images," *arXiv preprint*, 2022.
DOI: https://doi.org/10.48550/arXiv.2203.09777.
[9] L. Abady et al., "A Siamese-Based Verification System for Open-Set Attribution of Synthetic Images," *Pattern Recognition Letters*, 2024.
DOI: https://doi.org/10.1016/j.patrec.2024.02.012.
[10] M. Joslin et al., "Detecting and Attributing Fake Images Generated by GANs Using Frequency Analysis," *Digital Signal Processing Workshop*, 2020.
DOI: https://doi.org/10.1109/DSPW50219.2020.9282300.
[11] A. Saranya et al., "A Systematic Review of Explainable Artificial Intelligence Applications," *Machine Learning with Applications*, 2023.
DOI: https://doi.org/10.1016/j.mlwa.2023.100470.
[12] N. Bharati et al., "Explainable Deepfake Detection: A Multi-Model Framework," *Array*, 2025.
DOI: https://doi.org/10.1016/j.array.2025.100202.
[13] A. Mahara et al., "Methods and Trends in Detecting AI-Generated Images," *Computer Vision and Image Understanding*, 2026.
DOI: https://doi.org/10.1016/j.cviu.2026.103995.
[14] G. Nie et al., "Attributing Image Generative Models Using Latent Fingerprints," *Proceedings of Machine Learning Research*, 2023.
DOI: https://doi.org/10.48550/arXiv.2303.05050.
[15] B. Mohan Gurusamy et al., "Detecting AI-Generated Images with CNN and Interpretation Using Explainable AI," *IEEE International Conference on Contemporary Computing and Communications (InC4)*, 2024.
DOI: https://doi.org/10.1109/InC460750.2024.10649158