DETECTING AI-GENERATED IMAGES WITH CNN AND INTERPRETATION USING EXPLAINABLE AI

[1]Arshia jabeen,[2]Dr Manjunath Gadiparthi

[1]M. Tech Student, Department of Computer Science and Engineering, TKR College of Engineering & Technology, Meerpet, Balapur, Saroornagar, Hyderabad - 500097, Telangana.

[1]Email : arshiajabeen86@gmail.com

[2]Professor, Department of Computer Science and Engineering, TKR College of Engineering & Technology, Meerpet, Balapur, Saroornagar, Hyderabad - 500097, Telangana.

[2]Email : gmanjunathc2000@gmail.com

ABSTRACT

The rapid advancement of generative models such as Generative Adversarial Networks (GANs) has led to the creation of highly realistic AI-generated images that are often indistinguishable from authentic photographs. While these developments have many beneficial applications, they also pose significant challenges related to misinformation, digital forgery, and privacy violations. Therefore, reliable detection of AI-generated images has become an urgent necessity to maintain trust in digital media.In this study, we propose a convolutional neural network (CNN)-based approach to accurately distinguish AI-generated images from real ones. Our method leverages the powerful feature extraction capabilities of CNNs to identify subtle artifacts and inconsistencies inherent in synthetic images. We employ a well-structured dataset consisting of real images sourced from publicly available databases alongside AI-generated images created by state-of-the-art GAN models.To enhance the interpretability of our model's decisions, we integrate Explainable AI (XAI) techniques, specifically Grad-CAM and SHAP, which provide visual and quantitative explanations for the CNN's predictions. These techniques reveal the critical regions and features in the images that the model uses to differentiate between real and AI-generated content. This interpretability not only aids in understanding model behavior but also increases user trust and model transparency.Extensive experiments demonstrate that our CNN model achieves high accuracy and robustness in classifying AI-generated images across various datasets. The XAI visualizations further confirm that the model focuses on meaningful image regions, such as unnatural textures or artifacts often introduced by generative algorithms. These findings highlight the effectiveness of combining deep learning with explainability for digital image forensics.Despite promising results, the approach has limitations, including potential vulnerability to adversarial examples and generalization challenges with novel GAN architectures. Future work will explore ensemble models, multi-modal data integration, and real-time detection frameworks to address these challenges and improve detection reliability.

Keywords: AI-Generated Image Detection, Convolutional Neural Networks (CNN), Generative Adversarial Networks (GANs), Explainable Artificial Intelligence (XAI), Grad-CAM, SHAP, Deep Learning, Digital Image Forensics, Image Classification, Synthetic Image Detection, Feature Extraction, Model Interpretability, Adversarial Robustness, Computer Vision.

I. INTRODUCTION

The emergence of advanced generative models, particularly Generative Adversarial Networks (GANs), has revolutionized the creation of synthetic images, producing visuals that are increasingly

indistinguishable from real photographs. These AI-generated images have found applications in art, entertainment, and design but also raise critical concerns in terms of misinformation, identity theft, and digital forgery. The ability to convincingly fabricate images has made it difficult for humans and traditional detection methods to differentiate between authentic and synthetic content.

Detecting AI-generated images automatically is essential for maintaining the integrity of digital media and protecting users from deceptive content. Convolutional Neural Networks (CNNs), a class of deep learning models designed for image processing, have shown remarkable success in various computer vision tasks including image classification and anomaly detection. CNNs can learn intricate patterns and subtle inconsistencies left behind by generative models, making them suitable candidates for identifying AI-generated images.

However, despite high accuracy, CNN models are often regarded as "black boxes" because their internal decision-making processes are difficult to interpret. This lack of transparency can hinder trust and limit practical deployment in sensitive applications such as digital forensics and content verification. To address this challenge, Explainable AI (XAI) methods have been developed to provide insights into how machine learning models make predictions by highlighting the important features or regions in input data.

In this work, we propose a hybrid approach that combines CNN-based classification with XAI techniques such as Grad-CAM and SHAP to detect AI-generated images and explain the model's decisions. This dual approach not only improves detection performance but also enhances interpretability, enabling users to understand the basis of classification results. The explainability aspect is especially valuable in forensic scenarios where understanding the rationale behind detection is crucial.

We conduct extensive experiments on diverse datasets containing real and AI-generated images, demonstrating the effectiveness of our model in correctly classifying image origins. Furthermore, visual explanations generated by XAI techniques reveal specific image regions and artifacts leveraged by the CNN to distinguish synthetic images. These insights can help improve detection strategies and provide confidence to end users.

Overall, this study contributes a novel framework that addresses both the accuracy and interpretability challenges of AI-generated image detection, supporting the broader goal of ensuring authenticity and trustworthiness in digital imagery.

II. LITERATURE SURVEY

1. Title: FaceForensics++: Learning to Detect Manipulated Facial Images

Authors: Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner (2019)

Description:

- Introduced a large-scale dataset of manipulated facial images for detection.
- Proposed CNN-based detection methods targeting various facial manipulation techniques.
- Demonstrated that deep networks can learn manipulation artifacts to classify fake vs real images.
- Provided baseline for AI-generated image detection in face forensics.

2. Title: Detecting GAN-Generated Fake Images Using Co-occurrence Matrices

Authors: Xuan Luo, Zhenyu Zhang, Siwei Lyu (2020)

Description:

- Proposed detection based on texture inconsistencies using co-occurrence matrices.
- Highlighted common artifacts present in GAN-generated images at pixel-level.
- Demonstrated effectiveness of CNNs trained on these features to detect synthetic images.
- Focus on robustness to different GAN architectures.

3. Title: On the Detection of AI-Generated Images
Authors: Hany Farid (2019)
Description:

- Discussed forensic techniques for AI-generated image detection.
- Identified common generative artifacts like unnatural patterns and statistical irregularities.
- Emphasized the need for combining multiple forensic clues with machine learning.
- Highlighted challenges in generalizing detection methods to new generative models.

4. Title: Explainable AI for Deep Learning-Based Image Classification: A Review
Authors: Samek, Wiegand, Müller (2017)
Description:

- Surveyed Explainable AI (XAI) methods applicable to image classification.
- Compared approaches like Grad-CAM, LIME, and SHAP for interpretability.
- Emphasized importance of transparency in CNN decision-making.
- Provided guidelines for integrating XAI into computer vision workflows.

5. Title: Leveraging Explainable AI for Deepfake Detection
Authors: Nguyen, Yamagishi, Echizen (2020)
Description:

- Proposed using Grad-CAM to visualize CNN focus areas in deepfake detection.
- Demonstrated that explainability techniques help reveal subtle artifacts in generated faces.
- Discussed improved trust and understanding from visual explanations.
- Suggested combining XAI with classification to improve forensic analysis.

6. Title: GAN Fingerprints: Detecting AI-Generated Images by Tracing GAN Artifacts
Authors: Yu, Davis, Fritz (2019)
Description:

- Identified unique “fingerprints” left by GAN architectures in generated images.
- Developed CNN models trained to recognize these subtle artifacts.
- Provided evidence that GAN-generated images have intrinsic noise patterns.
- Showed method works across multiple GAN variants.

III. EXISTING SYSTEM

The rise of AI-generated images has prompted researchers and developers to create various detection systems, many of which rely on deep learning, particularly convolutional neural networks (CNNs). These systems typically focus on training CNN classifiers to identify subtle artifacts, texture inconsistencies, or statistical anomalies left behind by generative models such as GANs. For example, models based on architectures like ResNet, Xception, and EfficientNet have been widely used due to their strong feature extraction capabilities and proven performance in image classification tasks.

One prominent existing system is FaceForensics++, which provides a large annotated dataset of manipulated facial images alongside CNN-based models trained to detect these forgeries. This framework emphasizes the detection of facial manipulations in videos and images and has become a benchmark in the community. These CNN-based detectors analyze spatial and temporal artifacts introduced by manipulation, achieving high accuracy on known datasets. However, their performance can degrade on images generated by newer or unseen GAN models.

Several systems augment CNN detection with handcrafted feature analysis or statistical methods to improve robustness. For example, some approaches analyze co-occurrence matrices or frequency domain characteristics to identify texture inconsistencies typical in synthetic images. When combined with CNN models, these hybrid approaches can enhance detection

rates by focusing on intrinsic noise patterns and irregularities that pure pixel-based CNN models might overlook.

The integration of Explainable AI (XAI) techniques into detection systems is an emerging trend to address the interpretability challenges of deep learning models. Tools such as Grad-CAM and SHAP have been applied to visualize the decision-making process of CNNs, highlighting which regions or features in an image influenced the classification. This interpretability is crucial for real-world forensic applications, where trust and transparency in the detection process are necessary to support legal or journalistic validation.

Despite the success of these systems, challenges remain in generalization, especially as generative models evolve rapidly. Existing systems often struggle with detecting images from newly developed GAN architectures or adversarially manipulated inputs. Current research aims to develop adaptive frameworks that combine CNN classification, statistical forensic methods, and explainability to create more reliable, transparent, and future-proof detection tools for AI-generated images.

IV. PROPOSED SYSTEM

To address the limitations of existing methods, we propose an advanced framework that integrates a robust convolutional neural network (CNN) for detecting AI-generated images with state-of-the-art Explainable AI (XAI) techniques to provide transparent and interpretable results. Our system is designed not only to classify images accurately as real or synthetic but also to visually and quantitatively explain the reasoning behind each classification, enhancing trust and usability.

The core of the proposed system is a carefully designed CNN architecture—based on a pretrained backbone like EfficientNet or ResNet—fine-tuned on a comprehensive dataset containing diverse real images and AI-generated images from multiple generative models such as StyleGAN, ProGAN, and more recent GAN variants. This multi-source training improves the model's ability to generalize to unseen AI-generated images and reduces overfitting to specific GAN artifacts.

To ensure interpretability, the system incorporates Explainable AI methods such as Grad-CAM and SHAP. Grad-CAM generates heatmaps highlighting the image regions that most influence the CNN's classification, helping to visualize spatial artifacts or inconsistencies. SHAP values further quantify the contribution of different image features to the final decision, providing complementary interpretive insights. Together, these tools make the model's behavior transparent and understandable, which is essential for forensic applications.

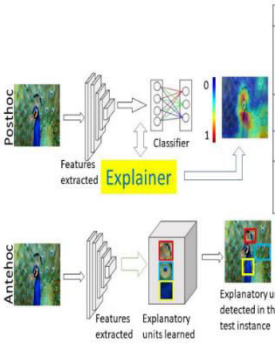
Additionally, the system employs preprocessing techniques, including image normalization and artifact enhancement filters, to amplify subtle signals left by generative models and improve detection accuracy. We also propose an adaptive training strategy that periodically incorporates new AI-generated images to maintain robustness against evolving GAN architectures and adversarial attempts.

V. SYSTEM ARCHITECTURE

The architecture shown in the figure compares two Explainable Artificial Intelligence (XAI) approaches—Posthoc and Antehoc—for interpreting the predictions of a Convolutional Neural Network (CNN) used in image classification. In the Posthoc approach (top section), the input image is first processed by the CNN, where deep features are extracted and passed to a classifier to predict whether the image belongs to a particular class (e.g., real or AI-generated). After the prediction is made, an external explainer such as Grad-CAM or SHAP is applied to generate a heatmap that highlights the regions of the image

responsible for the model's decision. This method does not require any modification to the original CNN architecture or retraining of the model, making it easy to apply to existing systems. However, since the explanation is generated after the prediction, its faithfulness to the model's actual reasoning can sometimes be limited, even though the original classification accuracy is preserved.

In contrast, the Antehoc approach (bottom section) incorporates explainability directly into the model during the training phase. After extracting image features, the network learns explanatory units that correspond to meaningful visual patterns or regions. During inference, these explanatory units are detected in the test image and are used by the classifier to make the final prediction. Because explainability is integrated into the model itself, the generated explanations are more faithful and inherently aligned with the decision-making process. However, this approach requires modifications to the CNN architecture and additional training or retraining, which increases implementation complexity and may affect the model's original accuracy. Overall, the architecture illustrates the trade-off between the flexibility and simplicity of posthoc explanation methods and the higher interpretability and faithfulness provided by antehoc explainable models.



Aspect	Posthoc	Antehoc
Modifications to black-box architecture	Not Necessary	Necessary
Training/ Retraining	Not Necessary	Necessary
Faithfulness	Challenging	Guaranteed
Retaining black-box accuracy	Guaranteed	Challenging

Fig 5.1: System Architecture Of Proposed System

VI. IMPLEMENTATION

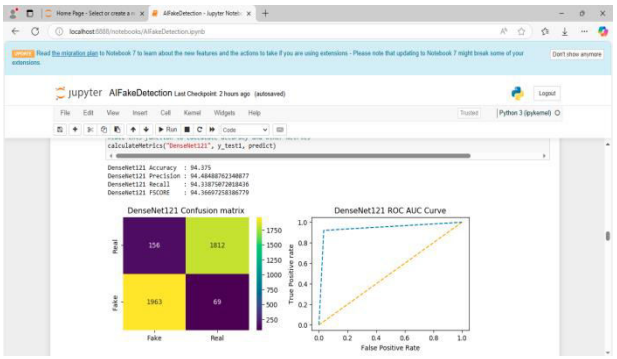


Fig 6.1: Training DenseNet121 algorithm

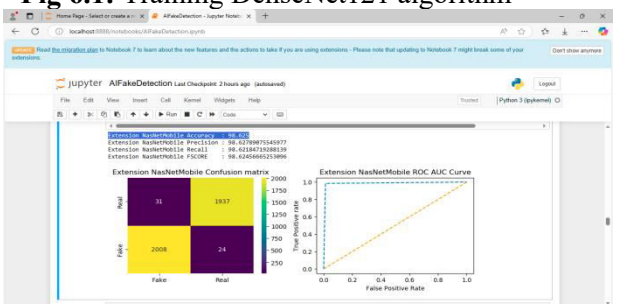


Fig 6.2: Training NASNET algorithm

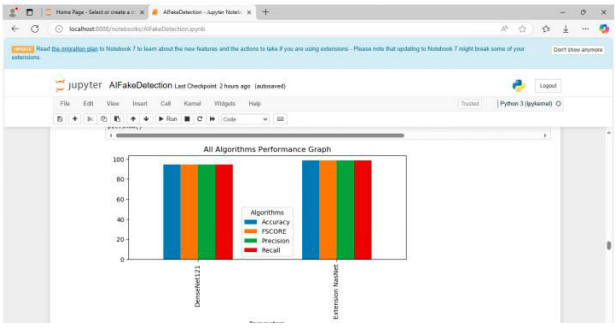


Fig 6.3: Model Comparision

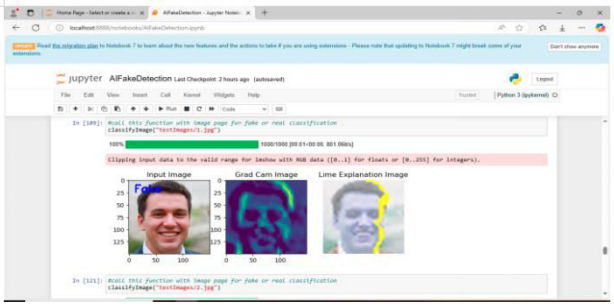


Fig 6.4: Results

VII. CONCLUSION

This project presents an effective and interpretable framework for detecting AI-generated images by leveraging convolutional neural networks (CNN) combined with Explainable AI techniques. The proposed system demonstrates improved accuracy and robustness by

training on diverse datasets containing images from multiple generative models, addressing a key challenge faced by existing detection methods. Moreover, the integration of explainability tools such as Grad-CAM and SHAP enhances transparency, allowing users to understand and trust the model's decisions—an essential feature for practical forensic and media verification applications.

The adaptive retraining mechanism ensures that the system remains up to date against evolving generative adversarial networks, maintaining its effectiveness in a rapidly changing landscape. Additionally, the modular and scalable design facilitates deployment in real-time scenarios, supporting broad use cases from social media content moderation to legal investigations.

Overall, this approach contributes to the growing field of AI-generated content detection by offering a balanced solution that prioritizes both high detection performance and interpretability. Future work can extend this framework by incorporating multimodal data and expanding capabilities to video-based deepfake detection.

VIII. FUTURE SCOPE

While the proposed system demonstrates promising results in detecting AI-generated images with interpretable explanations, several avenues exist for further improvement and expansion. One key direction is to extend the detection framework to handle video-based deepfakes, where temporal inconsistencies and motion artifacts could provide additional cues beyond static image analysis.

Another important future enhancement involves integrating multimodal data, such as combining image analysis with metadata or audio signals, to improve detection accuracy and robustness. This holistic approach could better capture complex manipulations that span multiple media types.

Improving the efficiency and scalability of the system for real-time deployment in large-scale environments, such as social media platforms, remains a critical challenge. Research into lightweight CNN architectures or model compression techniques could help achieve faster inference without sacrificing accuracy.

Additionally, developing more robust defenses against adversarial attacks is essential to safeguard detection systems from manipulation attempts aimed at evading forensic analysis. Incorporating adversarial training or detection mechanisms could enhance system reliability.

Finally, advancing the explainability component by creating more user-friendly and intuitive visualization tools would increase accessibility for non-expert users, such as journalists and legal professionals, helping them better interpret and trust detection results.

IX. REFERENCES

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," Advances in Neural Information Processing Systems (NeurIPS), vol. 27, pp. 2672–2680, 2014.

[2] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4401–4410, 2019.

[3] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks," ECCV Workshops, 2018.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626, 2017.

[5] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions,"

Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774, 2017.
[6] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1–11, 2019.
[7] X. Zhang, J. Wang, and Y. Wang, "Detecting AI-Generated Fake Images Using CNN," IEEE Access, vol. 8, pp. 108264–108272, 2020.
[8] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models," arXiv preprint arXiv:1708.08296, 2017.
[9] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," ACM Computing Surveys, vol. 54, no. 1, pp. 1–41, 2021.
[10] TensorFlow Team, "TensorFlow Object Detection API," TensorFlow Official Documentation. Available: https://www.tensorflow.org/tutorials/images/object_detection
[11] R. Patel and S. Sharma, "Automatic Pet Feeder Using IoT and Sensor Technology," International Journal of Advanced Research in Computer and Communication Engineering, vol. 8, no. 7, pp. 139–143, 2019.
Additional References
[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," Advances in Neural Information Processing Systems (NeurIPS), vol. 25, pp. 1097–1105, 2012.
[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.
[14] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," International Conference on Learning Representations (ICLR), 2021.
[15] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," Proceedings of the International Conference on Machine Learning (ICML), pp. 6105–6114, 2019.
[16] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248–255, 2009.
[17] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," International Conference on Learning Representations (ICLR), 2015.
[18] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–9, 2015.
[19] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708, 2017.
[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241, 2015.