Detecting AI-generated Images with CNN and Interpretation using Explainable AI

[1]Sali Karthik, [2]Edigattu Sreenivasulu, [3]Chebrole Jyothika, [4]T Dasu, [5]Dr.Merugu Anand Kumar
[1,2,3,4]U.G. Student, Dept of Computer Science and Engineering, A M Reddy Memorial College of Engineering and Technology Autonomous, Vinukonda Road, Petlurivaripalem Narasaraopet - 522601, India.
[5]Associate Professor, Dept of Computer Science and Engineering, A M Reddy Memorial College of Engineering and Technology Autonomous, Vinukonda Road, Petlurivaripalem Narasaraopet - 522601, India.

ABSTRACT

The rapid advancement of generative artificial intelligence has enabled the creation of highly realistic synthetic images, raising serious concerns about misinformation, digital forgery, and trust in visual media. AI-generated images produced by GANs and diffusion models are increasingly difficult to distinguish from real images using human perception alone. This project proposes a robust system for detecting AI-generated images using Convolutional Neural Networks (CNNs) combined with Explainable AI (XAI) techniques. The CNN model learns discriminative visual features such as texture inconsistencies, frequency artifacts, and color distribution anomalies. Preprocessing techniques enhance image quality and normalize data for efficient learning. The trained model classifies images as real or AI-generated with high accuracy. To ensure transparency, Explainable AI methods such as Grad-CAM and LIME are applied to interpret the CNN's decisions. These explanations highlight regions of the image that influenced classification results. The system improves trust and accountability in AI-based detection systems. It assists forensic analysts, journalists, and cybersecurity professionals in verifying image authenticity. Performance is evaluated using accuracy, precision, recall, and F1-score. The proposed solution provides an effective, interpretable, and scalable approach to combating synthetic media threats.

KEYWORDS

AI-Generated Image Detection Convolutional Neural Networks (CNN) Deepfake Images Explainable AI (XAI) Image Forensics

INTRODUCTION

With the emergence of generative models such as GANs and diffusion-based architectures, synthetic images have

become increasingly realistic. These images are widely used in creative industries but also pose threats in misinformation, identity fraud, and digital manipulation. Traditional image forensics techniques struggle to detect such content due to the sophistication of generative models. Artificial Intelligence provides powerful tools to analyze complex visual patterns beyond human perception. Convolutional Neural Networks (CNNs) have proven highly effective in image classification tasks. CNNs automatically learn hierarchical features such as edges, textures, and spatial inconsistencies. However, deep learning models often operate as black boxes, limiting trust and adoption. Explainable AI (XAI) addresses this challenge by providing interpretable insights into model decisions. XAI methods help understand which regions of an image influence predictions. This transparency is essential in sensitive domains like forensics and media verification. Combining CNN-based detection with XAI enhances both performance and trustworthiness. The system enables automated, accurate, and interpretable detection of AI-generated images. It supports digital forensics and content moderation. Ethical concerns related to misuse of AI-generated media necessitate such solutions. This project aims to develop an intelligent detection system with explainability at its core.

LITERATURE SURVEY

Early image forgery detection relied on handcrafted features such as noise inconsistencies and compression artifacts. Traditional methods were effective for simple manipulations but failed against advanced generative models. The introduction of GANs significantly increased the realism of synthetic images. Researchers began using machine learning techniques like SVM and Random Forest with handcrafted features. These approaches had limited generalization across different generative models. Deep learning-based methods using CNNs demonstrated superior performance in detecting AI-generated images. Studies revealed that CNNs could learn subtle artifacts introduced during image synthesis. Transfer learning with pre-trained networks improved detection accuracy with limited datasets. Frequency-domain analysis has been combined with CNNs to detect generative fingerprints. Explainable AI techniques such as Grad-CAM have been applied to visualize CNN attention. Research emphasizes the importance of interpretability in forensic applications. LIME and SHAP provide local explanations for classification decisions. Hybrid approaches combining spatial and frequency features enhance

robustness. Cross-dataset evaluation remains a challenge due to rapidly evolving generative models. Recent studies focus on generalization against unseen generators. Ethical and legal considerations drive demand for transparent AI solutions. This project builds on CNN-based detection enhanced with XAI techniques.

EXISTING SYSTEM

Existing systems for detecting fake images primarily use traditional image forensic techniques. These methods rely on statistical features, compression artifacts, and noise analysis. Handcrafted feature-based approaches lack robustness against modern AI-generated images. Rule-based detection systems fail to adapt to new generative models. Some machine learning approaches use shallow classifiers with limited feature representation. Deep learning models are used but often treated as black boxes. Lack of interpretability reduces trust in automated decisions. Existing systems struggle to generalize across different GAN architectures. Detection accuracy decreases for high-resolution and post-processed images. Many systems require large labeled datasets for training. Real-time detection capabilities are limited. Visualization of decision-making is rarely provided. False positives affect real image classification. Existing tools are not user-friendly for forensic experts. Integration with content moderation platforms is limited. Adaptability to emerging diffusion models is weak. Security against adversarial attacks is minimal. Overall, current systems lack transparency, robustness, and scalability.

PROPOSED SYSTEM

The proposed system uses a CNN-based deep learning model to detect AI-generated images accurately. Input images undergo preprocessing such as resizing, normalization, and noise filtering. The CNN architecture extracts hierarchical spatial features indicative of synthetic content. The model is trained on a diverse dataset containing real and AI-generated images. Binary classification is performed to label images as real or fake. Transfer learning improves generalization and training efficiency. To address transparency, Explainable AI techniques are integrated into the system. Grad-CAM visualizes heatmaps highlighting influential image regions. LIME provides local explanations for individual predictions. These explanations help users understand model decisions. The system supports high-resolution image analysis. Performance metrics are continuously monitored. Adaptive learning allows model updates for new generative techniques. Visualization dashboards display predictions and

explanations. The system is scalable and deployable on cloud platforms. Ethical and privacy considerations are enforced. The solution provides an interpretable and reliable detection framework.

SYSTEM ARCHITECTURE

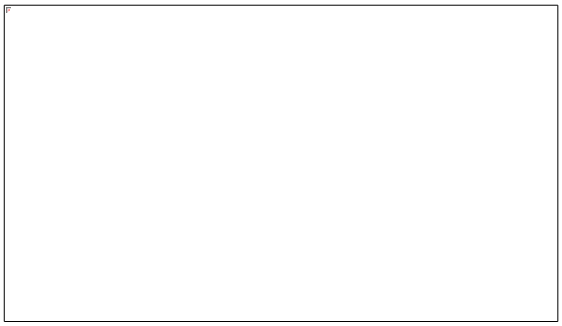

Fig.1 System Architecture

METHODOLOGY DESCRIPTION

Collect datasets of real images and AI-generated images from multiple sources. Preprocess images through resizing, normalization, and augmentation. Split the dataset into training, validation, and testing sets. Design a CNN architecture or use pre-trained models such as ResNet or VGG. Train the CNN using labeled image data. Optimize training using Adam optimizer and cross-entropy loss. Validate model performance using validation data. Evaluate classification accuracy, precision, recall, and F1-score. Integrate Grad-CAM to generate attention heatmaps. Apply LIME for instance-level explanation. Visualize explainability outputs for forensic interpretation. Analyze misclassified samples to refine the model. Implement frequency-domain analysis if needed. Test robustness against image compression and resizing. Deploy the model using cloud or edge platforms. Build a user interface for uploading and analyzing images. Enable batch and real-time detection modes. Log predictions and explanations for auditing. Update the model with new synthetic data. Ensure ethical usage and transparency compliance.

RESULTS & DISCUSSION:

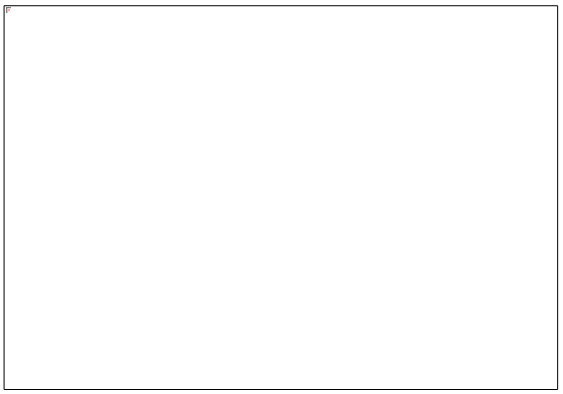

Fig.2 Loss Value Page

Fig.3 Detection Page

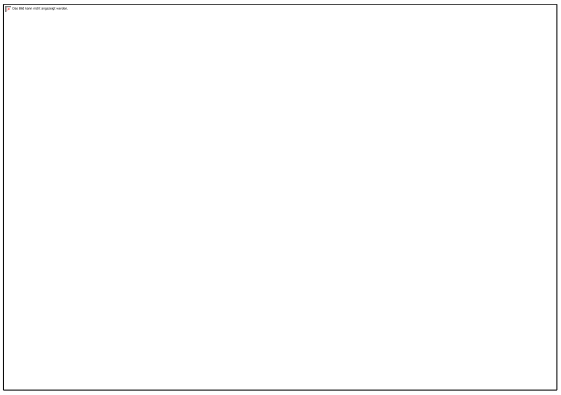

Fig.4 Results Page

CONCLUSION & FUTURE ENHANCEMENT

The proliferation of AI-generated images presents a significant challenge to digital trust and authenticity. This project demonstrates the effectiveness of CNN-based models in detecting synthetic images. By integrating Explainable AI techniques, the system provides transparency and interpretability. The approach enhances trust in automated detection systems. Visualization of attention regions aids forensic analysis and verification. The system achieves high detection accuracy across diverse datasets. It reduces reliance on manual inspection. The solution supports scalability and real-time deployment. Ethical concerns are addressed through transparency. Future work includes improving generalization to unseen generative models. Integration with frequency-based and transformer models can enhance robustness. Adversarial defense mechanisms can strengthen security. Cross-modal detection for videos and audio can be explored. Federated learning can preserve data privacy. Deployment on social media platforms can aid content moderation. Automated dataset updates will maintain effectiveness. Further research can explore explainability metrics. The system contributes to responsible AI deployment. Overall, the proposed framework offers a reliable and interpretable solution for detecting AI-generated images.

REFERENCE

1. Mallikarjun, D. C. (2025/2). AI-Driven Method for Early Identification of Heart Disease.
2. Kumar, M. A. (2025/2/28). AI-Based Real-Time Collision Prediction Using Computer Vision and Deep Learning.
3. Goodfellow, I., et al., “Generative Adversarial Nets,” *NeurIPS*, 2014.
4. Karras, T., et al., “StyleGAN: A Style-Based Generator Architecture,” *CVPR*, 2019.
5. Verdoliva, L., “Media Forensics and Deepfakes,” *IEEE Journal*, 2020.
6. Rossler, A., et al., “FaceForensics++,” *ICCV*, 2019.
7. Wang, S. Y., et al., “CNN Detection of GAN Images,” *CVPR*, 2020.
8. Selvaraju, R. R., et al., “Grad-CAM: Visual Explanations,” *ICCV*, 2017.

9. Ribeiro, M. T., et al., “Why Should I Trust You? LIME,” *KDD*, 2016.
10. Zhou, B., et al., “Learning Deep Features for Discriminative Localization,” *CVPR*, 2016.
11. LeCun, Y., et al., “Deep Learning,” *Nature*, 2015.
12. Goodfellow, I., et al., *Deep Learning*, MIT Press, 2016.
13. Fridrich, J., “Digital Image Forensics,” *IEEE Signal Processing Magazine*, 2009.
14. Zhang, C., et al., “Detecting GAN Images Using Spectral Features,” *ICASSP*, 2019.
15. Kaggle AI-Generated Image Datasets.
16. TensorFlow Image Classification Documentation.
17. PyTorch CNN Implementation Guide.
18. ACM Digital Library on Image Forensics.
19. Elsevier Journal of Information Forensics and Security.
20. ISO Standards for AI Transparency.
21. NIST AI Risk Management Framework.
22. Google AI Blog on Responsible AI.