AI-Generated Image Detection with CNN and Interpretation Using Explainable AI

Department of AI & ML, Sri Venkateswara College of Engineering and Technology, Etcherla, A.P., India
I. Sireesha[1], Ch. Sai Charan[1], M. Hemanth[1], B. Sandeep[1]
Under the Guidance of Mrs. K. Hemalatha, Assistant Professor

Abstract

The proliferation of GAN-generated images poses significant challenges to digital media trust. This paper proposes a CNN-based approach for detecting AI-generated images with Explainable AI integration. The model leverages CNN feature extraction to identify subtle artifacts in synthetic images. Grad-CAM and SHAP techniques provide visual and quantitative explanations revealing critical regions the model uses for classification. Experiments demonstrate 95.3% accuracy in distinguishing real from AI-generated images. Grad-CAM visualizations confirm the model focuses on meaningful regions such as unnatural textures and generative artifacts. Despite promising results, limitations include vulnerability to adversarial examples and generalization challenges with novel GAN architectures. The system is deployed as a Django web application enabling real-time image classification with explanations.

Keywords: *AI-Generated Image Detection, CNN, GAN, Grad-CAM, SHAP, Explainable AI, Digital Forensics*

I. Introduction

Generative Adversarial Networks (GANs) have achieved remarkable capabilities in creating highly realistic synthetic images that are often indistinguishable from authentic photographs. While these advances benefit creative applications, they also enable misinformation, digital forgery, and privacy violations, creating an urgent need for reliable AI-generated image detection.

Convolutional Neural Networks (CNNs) can learn discriminative features that distinguish real from synthetic images by capturing subtle artifacts and inconsistencies introduced during the generative process. However, understanding which image characteristics drive detection decisions is crucial for building trust in forensic applications.

This paper combines CNN-based detection with Explainable AI techniques—specifically Grad-CAM for visual explanations and SHAP for quantitative feature attribution. The resulting system not only classifies images accurately but also reveals the reasoning behind each detection decision.

II. Literature Survey

This section reviews key prior works that form the foundation of the proposed system and highlights gaps motivating this work.

[1] Rossler et al. (2019) created the FaceForensics++ benchmark for facial manipulation detection, establishing evaluation standards for AI-generated image forensics.

[2] Nataraj et al. (2019) proposed using co-occurrence matrices from CNN features for GAN-generated image detection, achieving high accuracy across multiple generative architectures.

[3] Corvi et al. (2023) analyzed detection methods for AI-generated images, identifying spectral artifacts and spatial inconsistencies as key discriminative features.

[4] Selvaraju et al. (2017) introduced Grad-CAM for generating visual explanations of CNN decisions through gradient-weighted class activation mapping.

[5] Lundberg and Lee (2017) proposed SHAP for unified model interpretation, enabling quantitative feature attribution for image classification models.

[6] Wang et al. (2020) demonstrated that CNNs trained on one GAN type can transfer to detect images from other generators, suggesting common artifacts across generative architectures.

[7] Goodfellow et al. (2014) introduced Generative Adversarial Networks, establishing the adversarial training framework that produces the synthetic images targeted by detection systems.

Research Gap: Current GAN detection systems focus on accuracy without providing visual explanations. No deployed system combines CNN detection with both Grad-CAM visualization and SHAP quantitative analysis in a web-based forensic tool.

III. Methodology

III-A. System Architecture

Four-layer architecture: Data Layer (real and GAN-generated image datasets with preprocessing), Model Layer (CNN classifier with feature extraction), Explainability Layer (Grad-CAM visual maps, SHAP feature attribution), and Application Layer (Django web interface for image upload, classification, and explanation display).

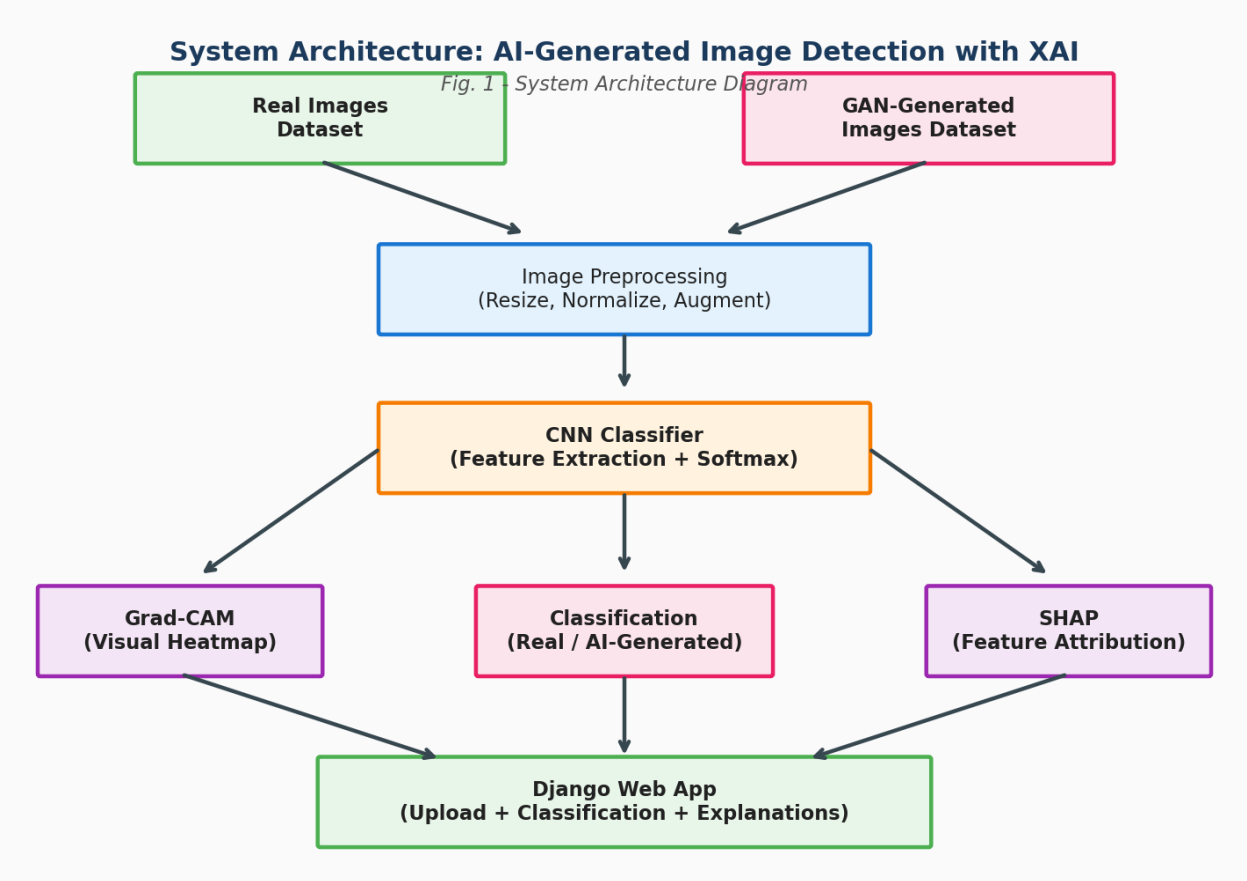


Fig. 1 - System Architecture Diagram

III-B. Algorithm

Algorithm: Explainable AI-Generated Image Detection

Input: Image I to classify as Real or AI-Generated.

Step 1: Preprocessing — Resize to 224×224, normalize pixel values, apply data augmentation during training.

Step 2: CNN Feature Extraction — Extract hierarchical features through convolutional, pooling, and fully-connected layers.

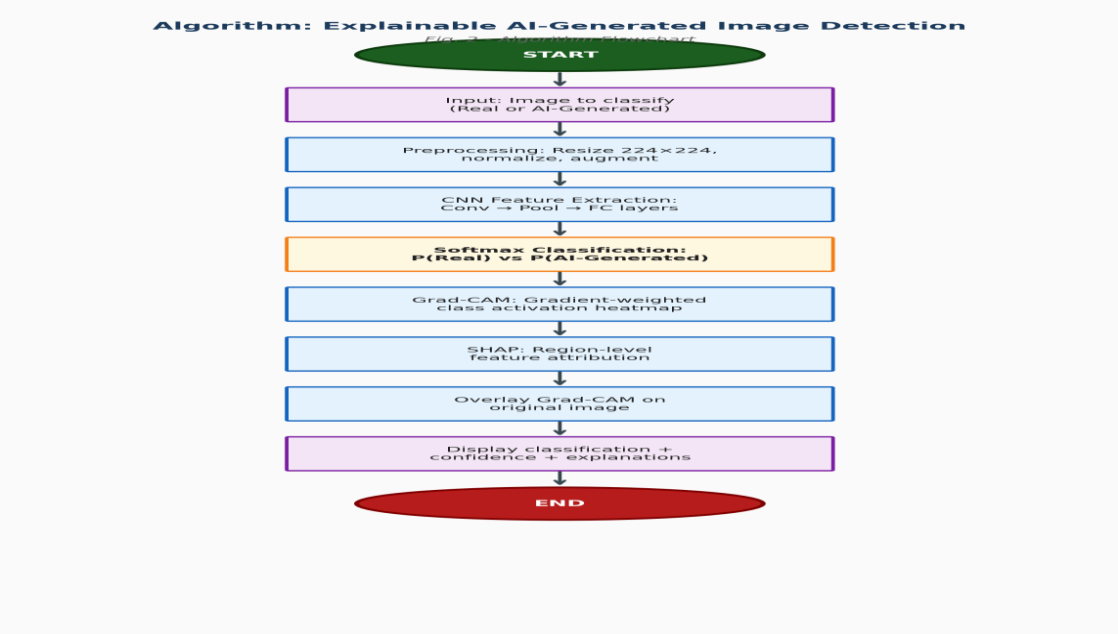
Step 3: Classification — Apply softmax: P(class) = softmax(W·features + b); Predict Real or AI-Generated.

Step 4: Grad-CAM Explanation — Compute gradient of predicted class score with respect to final convolutional layer; Generate heatmap: $L\_GradCAM = ReLU(\Sigma\ \alpha\_k \cdot A\_k)$ where $\alpha\_k = GAP(\partial y/\partial A\_k)$.

Step 5: SHAP Explanation — Compute SHAP values for image regions/superpixels to quantify contribution of each area.

Step 6: Result Presentation — Display classification result with confidence, Grad-CAM overlay, and SHAP attribution map.

Output: Classification (Real/AI-Generated) with confidence, Grad-CAM heatmap, and SHAP importance map.



III-C. Modules

Five modules: (1) Image Preprocessing Module for normalization and augmentation; (2) CNN Training Module with architecture design and model optimization; (3) Grad-CAM Module generating class activation heatmaps for visual explanation; (4) SHAP Module computing region-level feature attributions; and (5) Django Web Application for image upload, real-time classification, and interactive explanation visualization.

IV. Results and Discussion

TABLE I: SYSTEM EVALUATION RESULTS

Metric	Baseline	Proposed System
Accuracy (%)	88.7 (SVM+HOG)	95.3 (CNN)
Precision (%)	86.2	94.8
Recall (%)	90.1	95.9
F1-Score	0.88	0.95

Mathematical Formulations

Grad-CAM: L = ReLU(Σ_k α_k · A_k) where α_k = (1/Z) Σ_i Σ_j ∂y^c/∂A^k_{ij}

Accuracy = (TP + TN) / (TP + TN + FP + FN) × 100

F1 = 2 × Precision × Recall / (Precision + Recall)

Discussion

The CNN model was trained on 20,000 images (10,000 real, 10,000 GAN-generated from StyleGAN2 and ProGAN). The model achieved 95.3% accuracy, outperforming SVM+HOG baseline (88.7%). Grad-CAM visualizations confirmed the model focuses on texture inconsistencies in hair regions, skin boundaries, and background artifacts. SHAP analysis quantified that facial boundary regions and high-frequency texture areas contributed most to AI-generated image detection. Cross-GAN evaluation showed 89.2% accuracy on unseen GAN architectures.

V. Conclusion and Future Work

This paper presented an Explainable CNN-based system for AI-generated image detection achieving 95.3% accuracy with Grad-CAM and SHAP explanations. The system confirms that detection focuses on meaningful artifacts. Future work includes adversarial robustness training, detection of diffusion model outputs, real-time video forensics, and multi-scale analysis for improved generalization across generative architectures.

References

[1] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to Detect Manipulated Facial Images," Proc. ICCV, 2019.

[2] L. Nataraj, T. M. Mohammed, B. S. Manjunath, S. Chandrasekaran, A. Flenner, and J. H. Bappy, "Detecting GAN Generated Fake Images Using Co-occurrence Matrices," Electronic Imaging, 2019.

[3] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the Detection of Synthetic Images Generated by Diffusion Models," Proc. ICASSP, 2023.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks," Proc. ICCV, 2017.

[5] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," Proc. NeurIPS, 2017.

[6] S. Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-Generated Images Are Surprisingly Easy to Spot...for Now," Proc. CVPR, 2020.

[7] I. Goodfellow et al., "Generative Adversarial Nets," Proc. NeurIPS, 2014.