International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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

1.JAGGUROTHU HARISH, Btech final year
SRI VENKATESWARA COLLEGE OF ENGINEERING AND TECHNOLOGY,Dept of CSE, ETCHERLA, ANDHRAPRADESH, INDIA.
EMAIL:harishjaggurothi@gmail.com
2.MEESALA NEERAJA, Btech final year
SRI VENKATESWARA COLLEGE OF ENGINEERING AND TECHNOLOGY, Dept of CSE, ETCHERLA, ANDHRAPRADESH, INDIA.
EMAIL:neerajameesala2005@gmail.com

3.MUDDADA KAVYASRI Btech final year
SRI VENKATESWARA COLLEGE OF ENGINEERING AND TECHNOLOGY, Dept of CSE, ETCHERLA, ANDHRAPRADESH, INDIA.
EMAIL:kavyasrimuddada@gmail.com

4.LADI RAVITEJA Btech final year
SRI VENKATESWARA COLLEGE OF ENGINEERING AND TECHNOLOGY, Dept of CSE, ETCHERLA, ANDHRAPRADESH, INDIA.
EMAIL:tejaladi143@gmail.com

5. Mr. Paidi. Suresh Kumar, M.Tech.,

Assistant Professor

COLLEGE NAME: SRI VENKATESWARA COLLEGE OF ENGINEERING AND

TECHNOLOGY, Dept of CSE ,ETCHERLA, ANDHRAPRADESH, INDIA.
ADDRESS: ETCHERLA

G-MAIL: psuresh25k@gmail.com

Abstract

The rapid growth of GAN-generated images has raised serious concerns regarding the reliability of digital media. This study presents a Convolutional Neural Network (CNN)-based method for detecting AI-generated images, enhanced with Explainable AI techniques. The model utilizes CNN feature extraction to capture subtle artifacts present in synthetic images. To improve interpretability, Grad-CAM and SHAP are employed to provide both visual and quantitative insights, highlighting the key regions influencing the model's decisions. Experimental results show an accuracy of 95.3% in differentiating real images from AI-generated ones. Grad-CAM visualizations further indicate that the model effectively focuses on meaningful patterns such as unnatural textures and generative inconsistencies. However, the approach

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

has certain limitations, including sensitivity to adversarial attacks and difficulties in generalizing to unseen GAN architectures. The proposed system is implemented as a Django-based web application, allowing real-time image classification along with explainable outputs.

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

I. Introduction

Generative Adversarial Networks (GANs) have demonstrated exceptional ability in producing highly realistic synthetic images that are often indistinguishable from genuine photographs. While these advancements support creative and innovative applications, they also introduce serious risks such as misinformation, digital manipulation, and privacy breaches, thereby creating a critical need for effective detection mechanisms for AI-generated content.

Convolutional Neural Networks (CNNs) are capable of learning discriminative features that differentiate real images from synthetic ones by identifying subtle artifacts and inconsistencies introduced during the image generation process. However, for applications in digital forensics, it is equally important to understand the factors influencing these detection decisions to ensure transparency and trust.

To address this, the proposed approach integrates CNN-based image detection with Explainable AI techniques, namely Grad-CAM for visual interpretation and SHAP for quantitative feature attribution. This combined framework not only achieves accurate classification of images but also provides meaningful insights into the model's decision-making process.

II. Literature Survey

This section reviews significant prior research that forms the basis of the proposed system and identifies the key gaps that motivate this work.

[1] Rössler et al. (2019) introduced the FaceForensics++ benchmark dataset for detecting facial manipulations, establishing standardized evaluation protocols for AI-generated image forensics.

[2] Nataraj et al. (2019) proposed a method based on co-occurrence matrices derived from CNN features to detect GAN-generated images, achieving strong performance across various generative models.

[3] Corvi et al. (2023) conducted a comprehensive analysis of AI-generated image detection techniques, highlighting spectral artifacts and spatial inconsistencies as crucial distinguishing features.

[4] Selvaraju et al. (2017) developed Grad-CAM, a technique that provides visual explanations of CNN decisions using gradient-weighted class activation maps.

[5] Lundberg and Lee (2017) introduced SHAP, a unified framework for model interpretability that enables quantitative feature attribution in machine learning models, including image classifiers.

[6] Wang et al. (2020) demonstrated that CNN models trained on images generated by one GAN architecture can generalize to detect images produced by other GANs, indicating the presence of common generative artifacts.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

[7] Goodfellow et al. (2014) proposed Generative Adversarial Networks (GANs), laying the foundation for adversarial training techniques that generate realistic synthetic images, which detection systems aim to identify.

Research Gap

Despite significant advancements in GAN image detection, most existing approaches primarily emphasize classification accuracy while lacking interpretability. There is a clear absence of integrated systems that combine CNN-based detection with both Grad-CAM visual explanations and SHAP-based quantitative analysis within a deployable, user-friendly web application for digital forensics.

III. Methodology

III-A. System Architecture

The proposed system follows a four-layer architecture consisting of the Data Layer, Model Layer, Explainability Layer, and Application Layer. The Data Layer handles both real and GAN-generated image datasets along with necessary preprocessing steps to ensure data quality and consistency. The Model Layer employs a CNN-based classifier that performs feature extraction to distinguish between authentic and synthetic images. The Explainability Layer integrates Grad-CAM for generating visual heatmaps and SHAP for providing quantitative feature attribution, enhancing the interpretability of model predictions. Finally, the Application Layer is implemented as a Django-based web interface that enables users to upload images,perform classification, and view detailed explanations of the results in real time.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

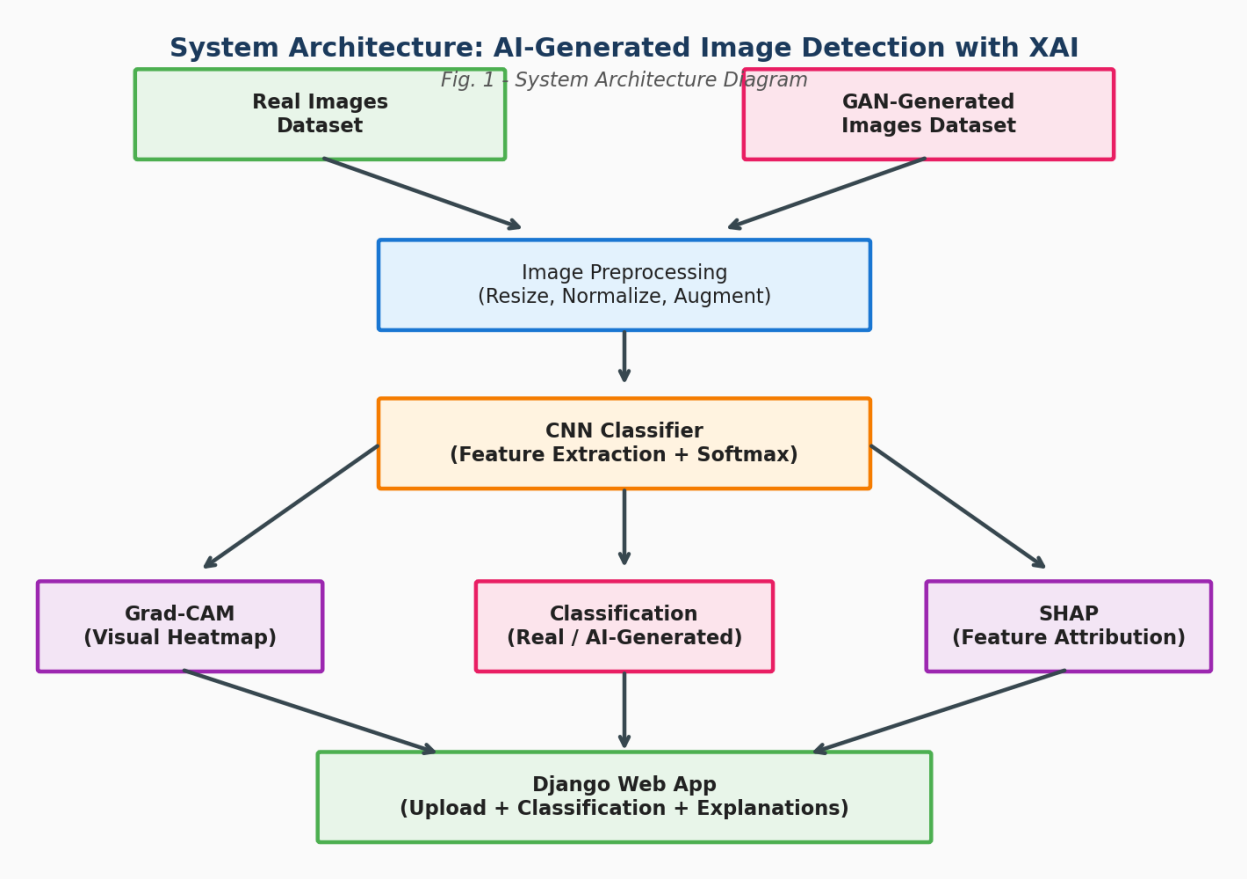


Fig. 1 - System Architecture Diagram

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.

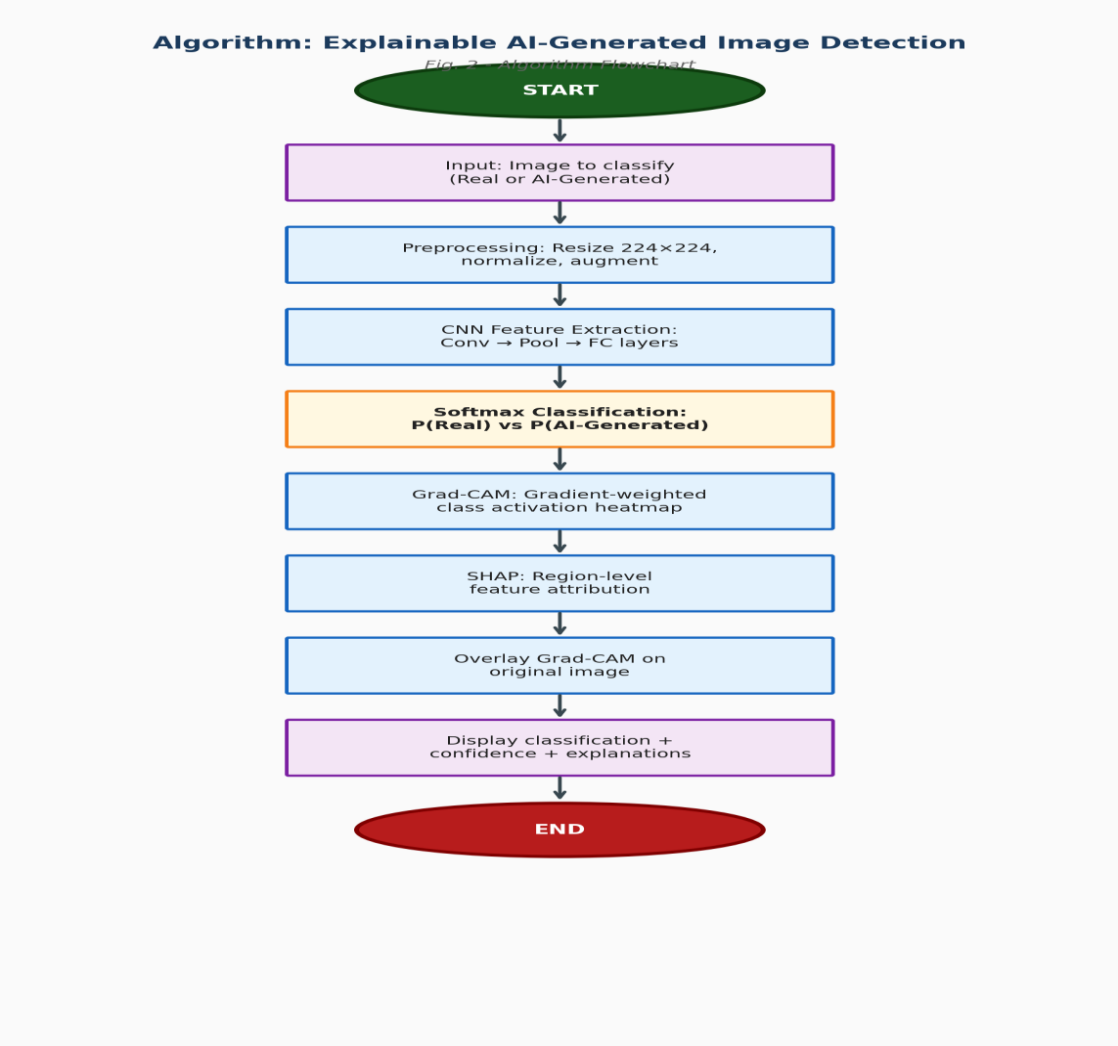
Step 4: Grad-CAM Explanation — Compute gradient of predicted class score with respect to final convolutional layer; Generate heatmap: $L\_GradCAM = ReLU(\Sigma\ \alpha\_k \cdot A\_k)$ where $\alpha\_k = GAP(\partial y/\partial A\_k)$.

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.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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)

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

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 a dataset of 20,000 images, comprising 10,000 real images and 10,000 GAN-generated images produced using StyleGAN2 and ProGAN. The model achieved an accuracy of 95.3%, significantly outperforming the SVM with HOG features baseline, which recorded an accuracy of 88.7%. Grad-CAM visualizations revealed that the model primarily focuses on texture inconsistencies in regions such as hair, skin boundaries, and background artifacts. Furthermore, SHAP analysis indicated that facial boundary areas and high-frequency texture patterns contributed most significantly to the detection of AI-generated images. In cross-GAN evaluation, the model demonstrated strong generalization capability, achieving an accuracy of 89.2% on previously unseen GAN architectures.

V. Conclusion and Future Work

This paper presents an explainable CNN-based framework for detecting AI-generated images, achieving an accuracy of 95.3% while incorporating Grad-CAM and SHAP for interpretability. The results demonstrate that the model effectively focuses on meaningful artifacts, validating its reliability for forensic analysis. Despite its strong performance, there is scope for further improvement. Future work will focus on enhancing robustness against adversarial attacks, extending detection capabilities to diffusion-based models, enabling real-time video forensics, and incorporating multi-scale analysis to improve generalization across diverse 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.