



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: <https://doi.org/10.22214/ijraset.2026.85278>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

An Explainable Deep Learning Framework for Detection of AI-Generated Images

Shitipragnya Sabat
Alard University, Pune

Abstract: *The rapid development of generative artificial intelligence has enabled the production of highly realistic synthetic images that are increasingly difficult to distinguish from authentic visual content. Although deep learning-based detection methods have demonstrated promising performance in identifying AI-generated images, many existing approaches operate as black-box models and provide limited explanations for their predictions. This lack of interpretability creates challenges for applications where users need to understand why an image has been classified as authentic or synthetic. This paper proposes an Explainable Deep Learning Framework for Detection of AI-Generated Images that combines deep visual feature learning with explainable artificial intelligence techniques. The proposed framework analyzes spatial and potentially frequency-domain characteristics of images, extracts discriminative features using a convolutional or transformer-based deep learning architecture, and classifies images into real and AI-generated categories. An explainability layer is incorporated using techniques such as Grad-CAM, saliency analysis, or related post-hoc explanation methods to identify image regions contributing to the model's decision. The framework is designed to evaluate not only classification performance but also robustness, generalization to unseen image-generation models, and interpretability. Existing literature indicates that cross-generator generalization remains a significant challenge because detector performance can decline when images originate from generators or scenes not represented during training. The proposed framework therefore emphasizes generalized detection and visual explanation rather than accuracy alone. The study is expected to contribute a transparent and reproducible approach for AI-generated image forensics and provide a foundation for trustworthy synthetic-media detection.*

Keywords: Artificial Intelligence, Deep Learning, AI-Generated Images, Synthetic Image Detection, Explainable AI, Computer Vision, Deepfake Detection, Image Forensics.

I. INTRODUCTION

Artificial intelligence has undergone substantial development in image generation, particularly through generative adversarial networks, diffusion models, and other generative architectures. These technologies can produce images with highly realistic visual characteristics and have applications in design, entertainment, education, advertising, research, and digital media. At the same time, the increasing realism of synthetic images creates challenges for determining whether a visual artifact originated from a camera or from an artificial generation system. AI-generated images can potentially be used for misinformation, identity manipulation, fraud, impersonation, copyright-related violations, and other forms of digital deception. Consequently, reliable identification of synthetic visual content has become an important research problem in multimedia security and digital image forensics. Recent surveys classify AI-generated image detection approaches into spatial-domain, frequency-domain, fingerprint-based, patch-based, and multimodal approaches, while also highlighting generalization and interpretability as important open problems.

Deep learning provides an effective approach for learning subtle patterns that distinguish authentic images from generated images. Unlike manually designed forensic features, deep neural networks can automatically learn discriminative representations from large image collections. However, high predictive performance does not necessarily mean that the model provides understandable reasoning. A classifier may correctly identify an image as synthetic while providing no information about which visual characteristics influenced the decision. This limitation is particularly important for forensic applications. A human analyst, researcher, journalist, platform administrator, or investigator may need evidence supporting an AI-generated classification rather than a simple binary output. Explainable Artificial Intelligence (XAI) provides techniques for interpreting model predictions and identifying features or image regions that contribute to a classification.

Recent research on explainable deepfake detection has examined techniques including Grad-CAM, LIME, SHAP, layer-wise relevance propagation, saliency maps, and related methods. This research direction emphasizes the need to connect detection accuracy with transparency and interpretability.

Therefore, this research proposes an explainable deep learning framework for AI-generated image detection. The central objective is not merely to determine whether an image is real or synthetic, but to provide a visual explanation of the model's decision.

II. PROBLEM STATEMENT

The increasing quality of AI-generated images has reduced the reliability of conventional visual inspection. Human observers may be unable to consistently distinguish synthetic images from authentic photographs, particularly when generated images undergo resizing, compression, cropping, filtering, or other post-processing operations.

Deep learning-based detectors can learn subtle artifacts that may not be visible to humans. However, several challenges remain:

- 1) Detection models may learn generator-specific artifacts rather than general characteristics of synthetic images.
- 2) Performance can decrease when testing images generated by previously unseen models.
- 3) Image compression and post-processing can remove or modify forensic evidence.
- 4) Many deep learning models provide predictions without understandable explanations.
- 5) High classification accuracy alone does not establish the reliability or practical usefulness of a forensic detector.
- 6) A detector trained on one type of synthetic image may not generalize to newer generation technologies. Research has specifically reported performance degradation across generators and image domains, demonstrating that cross-domain generalization remains a major challenge.

Hence, there is a need for a framework that simultaneously considers detection, generalization, robustness, and explain ability.

III. RESEARCH GAP

Existing research demonstrates substantial progress in AI-generated image detection. Nevertheless, the following research gaps can be identified:

- 1) Limited Cross-Generator Generalization: Many detection systems perform well on images generated by models represented in the training dataset but may experience reduced performance when evaluated on unseen generators or different image domains.
- 2) Accuracy-Centric Evaluation: A large proportion of detection studies emphasize accuracy, precision, recall, F1-score, or AUC. Although these metrics are essential, they do not explain why a model makes a particular decision.
- 3) Limited Integration of Explainability: Explainable AI is increasingly being studied in synthetic-media detection, but the integration of visual explanations into an end-to-end AI-generated image detection framework remains an important area for investigation. Recent literature specifically identifies explainability as a major challenge in deepfake detection.
- 4) Robustness to Post-Processing: Images shared through social media and digital platforms are commonly compressed, resized, cropped, or filtered. Such transformations can modify forensic traces and reduce detector performance.
- 5) Difficulty Detecting New Generation Models: Generative image technology continues to evolve. A detector developed for older GAN-generated images may not necessarily perform equally well on images produced by newer diffusion-based systems. Recent research has consequently explored frequency information, reconstruction errors, artifact purification, and other features for improved generalization.

IV. RESEARCH AIM

The primary aim of this research is:

To design and evaluate an explainable deep learning framework capable of detecting AI-generated images while providing interpretable visual evidence supporting its classification decisions.

V. RESEARCH OBJECTIVES

The study has the following objectives:

- 1) To investigate existing deep learning techniques for AI-generated image detection.
- 2) To construct a suitable dataset containing authentic and AI-generated images.
- 3) To develop a deep learning model for binary classification of real and AI-generated images.
- 4) To investigate spatial and frequency-domain features relevant to synthetic image detection.
- 5) To integrate an Explainable AI technique with the proposed detection model.
- 6) To visualize image regions that influence the model's classification.
- 7) To evaluate the model using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix.
- 8) To investigate model robustness against image compression, resizing, cropping, and other transformations.
- 9) To evaluate cross-generator generalization using unseen AI image generators.
- 10) To compare the proposed framework with selected existing deep learning approaches.

VI. RESEARCH QUESTIONS

The study seeks to answer the following research questions:

- 1) RQ1: How effectively can deep learning distinguish AI-generated images from authentic images?
- 2) RQ2: Which visual or frequency-domain characteristics contribute most strongly to AI-generated image detection?
- 3) RQ3: Can an explainability mechanism provide meaningful visual evidence for the detector's predictions?
- 4) RQ4: How well does the proposed framework generalize to AI-generated images produced by previously unseen generation models?
- 5) RQ5: How does image post-processing affect the performance of the proposed detection framework?
- 6) RQ6: What is the relationship between detection performance and explanation quality?

VII. PROPOSED METHODOLOGY

The proposed methodology consists of the following major stages:

Dataset Collection → Image Pre-processing → Feature Representation → Deep Learning Model → Classification → Explainability Layer → Robustness Testing → Performance Evaluation

A. Dataset Collection

The dataset should contain two principal classes:

- Class 0: Authentic Images
- Class 1: AI-Generated Images

Authentic images may be collected from publicly available image datasets with appropriate research usage permissions. AI-generated images should preferably represent multiple generation approaches rather than a single generator.

The dataset can include images generated using different categories of generative models, such as:

- GAN-based generators
- Diffusion-based generators
- Text-to-image systems
- Other publicly available synthetic-image sources

A multi-generator dataset is preferable because recent research demonstrates that detector performance may decline when evaluated across different generators and domains.

B. Dataset Pre-processing

Before training, images should undergo standardized preprocessing:

- 1) Removal of corrupted files.
- 2) Duplicate detection.
- 3) Image resizing.
- 4) Pixel normalization.
- 5) Conversion into a consistent image format.
- 6) Optional data augmentation.
- 7) Training-validation-test splitting.

For example:

Dataset	Suggested Proportion
Training	70%
Validation	15%
Testing	15%

Care must be taken to prevent near-duplicate images from appearing across different subsets.

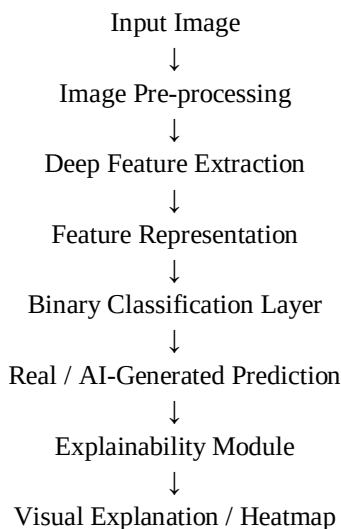
VIII. PROPOSED DEEP LEARNING ARCHITECTURE

The proposed framework can employ a modern image classification architecture such as:

- ResNet
- EfficientNet
- ConvNeXt
- Vision Transformer
- Swin Transformer

For an initial implementation, **EfficientNet or ResNet** can provide a practical baseline because they offer established transfer-learning workflows and comparatively manageable computational requirements.

The architecture can be represented as:



The output can be represented as:

Prediction: AI-Generated

Confidence: XX%

Important Regions: Highlighted Heatmap

The confidence value should be generated by the trained model and should not be interpreted as an absolute probability of truth without calibration.

IX. EXPLAINABLE AI MODULE

The major contribution of the proposed framework is the incorporation of an explainability layer.

A. Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) can be used to identify image regions that contribute to the model's prediction.

The general process is:

1. Input an image into the trained model.
2. Obtain the predicted class.
3. Calculate gradients associated with the target class.
4. Determine the importance of feature maps.
5. Generate an activation heatmap.
6. Overlay the heatmap on the original image.

The resulting visualization can help determine whether the model focuses on meaningful forensic regions or irrelevant image areas.

B. Alternative XAI Methods

Additional methods may be evaluated:

- LIME
- SHAP
- Integrated Gradients
- Saliency Maps
- Layer-wise Relevance Propagation

Recent research on explainable deepfake detection identifies several of these techniques as important approaches for improving transparency in synthetic-media detection.

C. Explanation Evaluation

The study should not simply generate heatmaps. It should investigate whether the explanations are meaningful.

Possible evaluation criteria include:

- Localization consistency
- Explanation stability
- Sensitivity to image perturbations
- Human interpretability
- Faithfulness to model predictions

This transforms explainability from a visualization feature into an evaluable research component.

X. FREQUENCY-DOMAIN ANALYSIS

In addition to spatial image features, the framework can investigate frequency-domain information.

Synthetic image generation may introduce patterns or inconsistencies that are not obvious in the RGB representation. Frequency-based methods analyze information using transformations such as Fourier or discrete cosine transforms.

Recent research has demonstrated the usefulness of frequency-related information for detecting diffusion-generated images. For example, FIRE investigates frequency-guided reconstruction error for improved detection and generalization to unseen diffusion models.

A possible extension of the proposed framework is therefore:

RGB Features + Frequency Features → Feature Fusion → Deep Classifier

This hybrid design could potentially improve robustness compared with relying exclusively on spatial features.

XI. PROPOSED FRAMEWORK

The complete framework is proposed as follows:

Stage 1: Input

The system receives a digital image.

Stage 2: Pre-processing

The image is resized and normalized.

Stage 3: Feature Extraction

Deep learning extracts spatial features.

Optional frequency-domain processing extracts additional forensic characteristics.

Stage 4: Feature Fusion

Spatial and frequency representations are combined.

Stage 5: Classification

The classification model predicts:

- Real Image
- AI-Generated Image

Stage 6: Explainability

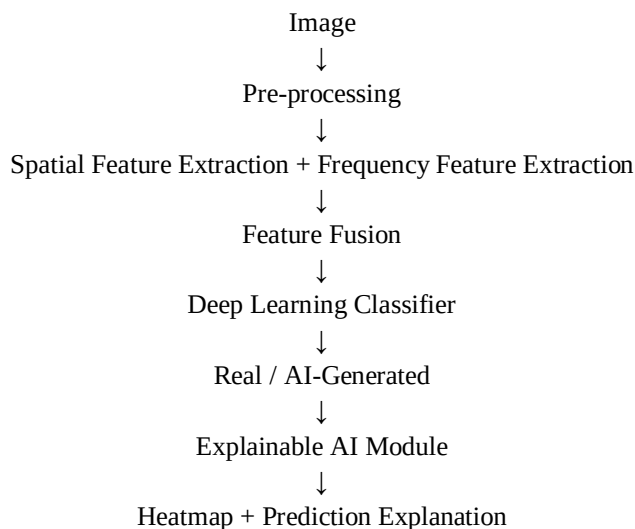
The XAI module identifies the image regions contributing to the prediction.

Stage 7: Output

The final system produces:

- Predicted class
- Confidence score
- Explanation heatmap
- Optional forensic feature summary

XII. CONCEPTUAL FRAMEWORK



XIII. EXPERIMENTAL DESIGN

The experimental study should compare multiple model configurations.

Experiment 1: Baseline Model

Train a conventional deep learning classifier using RGB images.

Experiment 2: Explainable Model

Add the XAI component to the baseline model.

Experiment 3: Frequency-Enhanced Model

Combine spatial and frequency-domain information.

Experiment 4: Cross-Generator Testing

Train using selected generators and test using a generator not included in training.

Experiment 5: Robustness Testing

Apply transformations such as:

- JPEG compression
- Image resizing
- Cropping
- Blur
- Noise
- Brightness modification
- Contrast modification

Then evaluate whether the model continues to identify synthetic images correctly.

XIV. PERFORMANCE EVALUATION METRICS

The following metrics should be used:

Accuracy

Measures the overall proportion of correctly classified images.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Precision

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-Score

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC-AUC

ROC-AUC can be used to evaluate the model's discrimination capability across different classification thresholds.

Confusion Matrix

The confusion matrix should report:

- True Positive
- True Negative
- False Positive
- False Negative

For this research, evaluation should extend beyond accuracy because a detector that performs well on a familiar generator but poorly on unseen generators may have limited practical forensic value.

XV. PROPOSED RESULTS PRESENTATION

Important: The following table is a results template. Actual values must be inserted after conducting the experiments.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
ResNet Baseline	—	—	—	—	—
EfficientNet Baseline	—	—	—	—	—
Proposed Explainable Model	—	—	—	—	—
Proposed Frequency-Enhanced Model	—	—	—	—	—

Cross-Generator Evaluation

Training Generator(s)	Testing Generator	Accuracy	F1-Score
Generator A	Generator A	—	—
Generator A	Generator B	—	—
Generator A + B	Generator C	—	—
Multiple Generators	Unseen Generator	—	—

Robustness Evaluation

Transformation	Accuracy	F1-Score
Original	—	—
JPEG Compression	—	—
Resizing	—	—
Cropping	—	—
Gaussian Noise	—	—
Blur	—	—

XVI. EXPLAINABILITY EVALUATION

The explainability component should be evaluated separately from classification performance.

For correctly classified AI-generated images, Grad-CAM or another selected XAI technique can generate heatmaps.

The analysis should determine whether the model focuses on:

- Facial regions
- Object boundaries
- Textures
- Background inconsistencies
- Repeated patterns
- Fine-grained image artifacts
- Other forensic regions

An important research question is whether the model is identifying genuine synthetic-image evidence or simply relying on dataset-specific characteristics.

For example, if the model consistently highlights irrelevant backgrounds rather than forensic regions, this may indicate dataset bias.

XVII. EXPECTED CONTRIBUTIONS

The proposed research is expected to make the following contributions:

Contribution 1: Explainable Detection Framework

A framework that combines deep learning-based classification with visual explanations for AI-generated image detection.

Contribution 2: Multi-Domain Feature Analysis

Investigation of spatial and frequency-domain characteristics for identifying synthetic images.

Contribution 3: Cross-Generator Evaluation

Evaluation of model performance on generators that were not included in training.

Contribution 4: Robustness Analysis

Systematic evaluation under common image transformations and post-processing operations.

Contribution 5: Transparent Decision Support

Generation of visual evidence explaining the regions that influence the detector's classification.

XVIII. DISCUSSION

The proposed framework addresses an important limitation in AI-generated image detection: the difference between prediction and understandable evidence.

Existing research demonstrates that the detection problem is becoming more difficult as generative models improve. Current literature also emphasizes the importance of cross-model generalization because detector performance can decline when the generation model or image domain changes.

Recent approaches have investigated frequency characteristics, reconstruction errors, artifact purification, and specialized architectures to improve robustness against unseen generators.

The proposed research adds another dimension by explicitly incorporating explainability. This is important because a binary prediction such as “AI-generated” does not provide sufficient information for many forensic and decision-support scenarios.

The proposed framework therefore evaluates three related dimensions:

Detection Performance + Generalization + Explainability

rather than treating classification accuracy as the only measure of success.

XIX. LIMITATIONS

The proposed study may have several limitations.

First, the quality and diversity of the dataset can strongly affect model performance. A dataset containing only a small number of generators may produce misleadingly high results.

Second, new image-generation models may introduce artifacts that are not represented in the training data.

Third, explainability techniques such as Grad-CAM provide post-hoc interpretations and should not automatically be considered a complete representation of the model's internal reasoning. Research on interpretable deep models highlights the distinction between post-hoc explanations and models designed to be interpretable by construction.

Fourth, image compression, resizing, filtering, and other transformations may alter forensic evidence.

Finally, high classification accuracy on a benchmark dataset does not guarantee real-world performance.

XX. FUTURE SCOPE

Future research can extend the proposed framework in several directions.

19.1 Multimodal Synthetic Media Detection

The framework can be extended from images to combinations of:

- Image
- Video
- Audio
- Text

Recent research increasingly treats synthetic-media detection as a multimodal problem.

19.2 Vision-Language Models

Vision-language models could be investigated for combining visual evidence with semantic reasoning.

19.3 Continual Learning

The detector could continuously learn from newly emerging image-generation technologies.

19.4 Lightweight Detection

A compressed model could be developed for deployment on mobile devices, browsers, or edge systems.

19.5 Content Provenance

Detection could be combined with watermarking, content provenance, and authentication mechanisms. NIST identifies detection, labeling, provenance, watermarking, and auditing among the technical approaches relevant to synthetic-content transparency.

19.6 Human-AI Collaboration

Future systems could provide explanations to human forensic analysts rather than replacing human decision-making.

XXI. CONCLUSION

This paper proposed an Explainable Deep Learning Framework for Detection of AI-Generated Images. The framework addresses the growing challenge of distinguishing synthetic images from authentic visual content while recognizing that reliable detection requires more than classification accuracy.

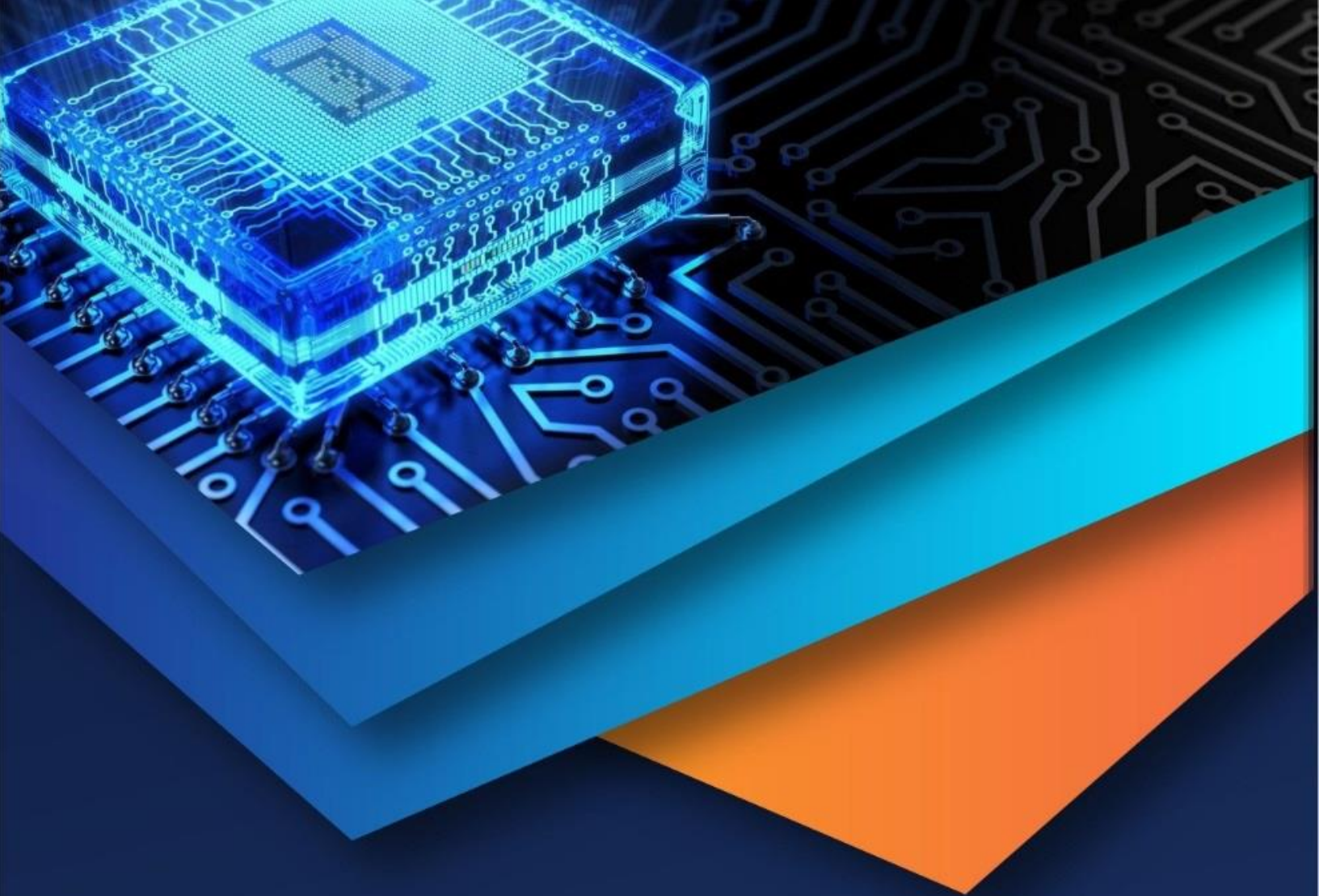
The proposed approach combines deep learning-based visual feature extraction, optional frequency-domain analysis, binary classification, cross-generator evaluation, robustness testing, and an Explainable AI layer. The explainability component is intended to provide visual evidence showing which regions contribute to the model's prediction.

The research is particularly focused on addressing the challenges of generalization, robustness, and transparency. Existing literature indicates that detector performance can decline across unseen generators and domains, while explainability remains an important requirement for trustworthy synthetic-media detection.

The proposed framework can therefore provide a foundation for developing transparent AI-generated image forensic systems. Its final effectiveness should be established through controlled experiments using diverse datasets, unseen generators, post-processing transformations, and quantitative as well as explainability-oriented evaluation.

REFERENCES

- [1] Zhang, Y., Pang, Z., Huang, S., et al. (2025). Unmasking AI-created visual content: A review of generated images and deepfake detection technologies. *Journal of King Saud University Computer and Information Sciences*, 37, 148.
- [2] Meng, Z., Peng, B., Dong, J., Tan, T., & Cheng, H. (2024). Artifact feature purification for cross-domain detection of AI-generated images. *Computer Vision and Image Understanding*, 247, 104078.
- [3] Chu, B., Xu, X., Wang, X., Zhang, Y., You, W., & Zhou, L. (2025). FIRE: Robust Detection of Diffusion-Generated Images via Frequency-Guided Reconstruction Error. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2025)*, 12830–12839.
- [4] Deng, J., Lin, C., Zhao, Z., et al. (2025). A Survey of Defenses Against AI-Generated Visual Media: Detection, Disruption, and Authentication. *ACM Computing Surveys*, 58(5), Article 123.
- [5] Explainable deepfake detection across different modalities: An overview of methods and challenges. (2025). *Image and Vision Computing*, 163, 105738.
- [6] Tan, D., Niu, C., Yang, Y., Yang, D., & Tan, B. (2025). DC-BiNet: Towards interpretable generated image detection with dark channel prior. *Expert Systems with Applications*, 280, 127508.
- [7] Wu, Y., Zhang, F., Shi, T., et al. (2025). Explainable Synthetic Image Detection through Diffusion Timestep Ensembling.
- [8] Di Marino, A., Bevilacqua, V., Ciaramella, A., De Falco, I., & Sannino, G. (2025). Ante-Hoc Methods for Interpretable Deep Models: A Survey. *ACM Computing Surveys*, 57(10), Article 262.
- [9] Chandra, B., Dunietz, J., Roberts, K., Lee, Y., Fontana, P., & Awad, G. (2024). Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency. *NIST Trustworthy and Responsible AI, NIST AI 100-4*.
- [10] Survey Paper: Comparative Analysis of Detection Methods for Real vs. AI-Generated Images. (2025). *2025 International Conference on Artificial Intelligence and Data Engineering (AIDE)*. IEEE.
- [11] Methods and trends in detecting AI-generated images: A comprehensive review. (2026). *Computer Science Review*.
- [12] A Survey on Deep-Learning-Based Techniques for Detecting AI-Generated Synthetic Images. (2026). *Engineering Proceedings*, 123(1), 32.



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)