AI-Generated Artwork Detection Using Error Level Analysis and a Hybrid Deep Learning Model with Explainable AI - A Review

Ms. Trupti Deore[1], Vaibhav Dattu Pawar[2], Namrata Suresh Bidgar[3], Kajal Dattu Rathod[4]

Department of AI & Data Science, MCOERC, Nashik, India

Abstract

Image generators based on deep learning now turn out digital paintings that even trained viewers struggle to separate from hand-made work. Verifying the origin of a picture has consequently become a pressing need in art markets, galleries, museums and web platforms. This paper surveys recent research on identifying machine-made artwork, grouping it into conventional classifiers, convolutional networks, transformer-based detectors and explanation techniques, and notes which problems are still open. Guided by that survey, a detection pipeline is outlined in which Error Level Analysis (ELA) supplies compression-based forensic cues and a hybrid network, pairing a convolutional branch for local texture with a transformer branch for image-wide context, decides whether a piece is AI-generated or human-created. Training and testing are planned on the AI-ArtBench collection, and Grad-CAM heat maps will show the image regions behind each decision. The paper further covers the architecture, the planned evaluation protocol and the known limits of the approach. The intended outcome is a dependable and transparent tool that artists, buyers, curators, researchers and regulators can use to check digital artwork.

Keywords: AI-generated artwork, error level analysis, hybrid deep learning, explainable AI, Grad-CAM, AI-ArtBench, image forensics

I. INTRODUCTION

Image-generating software has changed how pictures are made. A short text prompt is now enough to obtain a painting, an illustration or a lifelike design that used to take a skilled artist many hours. Most of this progress rests on two model families: generative adversarial networks [1] and diffusion models [2], [3]. Alongside the creative gains, these tools raise doubts about who the author of a work is, how original it is and whether it can be considered genuine.

The worry is a practical one for working artists. A generated image can be presented as hand-made, sold in the same marketplaces, or built to mimic the style of a living painter. Buyers, exhibition spaces and web platforms must in turn decide whether a piece is what it claims to be. The traces left by the generation process are tiny and escape the human eye, so visual inspection by people loses reliability as the images get more lifelike [4].

Automated detection has therefore become a busy research topic. Initial studies trained conventional classifiers and convolutional neural networks (CNNs) to tell machine-made paintings from human ones [5], [6], while newer studies adopt transformer models paired with visual explanations [7], [8]. Still, a detector

tuned on one generator or one style can lose accuracy on another, and a bare "AI" or "human" verdict offers the user little basis for confidence.

The paper has two aims. It first examines recent work on detecting machine-made artwork and identifies what each family of methods fails to address. It then proposes a framework that pairs Error Level Analysis (ELA) [9], used to pull out forensic cues, with a hybrid CNN–transformer classifier and with explainable AI that visually backs each prediction.

The remainder is arranged as follows. Section II surveys earlier studies and their shortcomings, Section III describes the proposed design, Section IV sets out the evaluation plan and expected outcomes, and Section V closes the paper.

II. LITERATURE REVIEW

Work on distinguishing AI-generated images from human-made ones has grown quickly since diffusion models became widely available. This section groups the important studies into five themes and then lists the gaps that remain.

A. Classical Machine Learning and CNN-Based Detection

The first line of work treats detection as an ordinary image-classification task. Li and Stamp [5] tested Logistic Regression, Support Vector Machine, Multilayer Perceptron and CNN models on human artwork from WikiArt and AI artwork from AI-ArtBench, in three styles. Their best results were 97.58% accuracy for the binary AI-versus-human problem, while the harder six-class problem (style plus origin) reached 82.08%. Chinta et al. [6] compared a custom CNN, VGG-19 and ResNet-50 [10] on AI-ArtBench, and the custom CNN gave the best accuracy of 92.69%.

Other studies in this group include a CNN-based separation of AI-generated art from human art by Tinago et al. [11], the benchmarking of several pretrained CNN backbones on the same task by Alrawahneh et al. [12], and the "AI vs. AI" study of Baraheem and Nguyen [13], which asks whether AI models can detect AI-generated images. Together they show that standard CNNs already work well on the binary task.

Their weakness is generalisation. These models learn from the pixels of a particular dataset and may fail when the images come from a new generator or an unseen artistic style, a limitation that the authors of [5] also point to when they suggest testing newer generative models.

B. Transformer-Based and Attention-Based Detectors

A second group uses attention and transformer architectures, which model relations between distant parts of an image and not only local textures. Vision transformers [14] and modern CNN designs such as ConvNeXt [15] are common backbones. Wang et al. [7] proposed a self-distilled transformer built on DINOv2 [16] that learns both global and local features, and reported 99.01% accuracy with 95.29% precision, 94.58% recall and 94.93% F1-score. Silva et al. [8] built ArtBrain, an end-to-end toolkit based on an AttentionConvNeXt network, which classifies AI-generated art, identifies the generating model and the artistic style, and reached about 99% accuracy in identifying the generative model.

These results are strong, but both papers list the same open problem: keeping up with newer image generators and larger, more diverse datasets. Neither is reported to use forensic image features as an explicit input.

C. Forensic Features and Error Level Analysis

Digital image forensics looks for traces that are invisible to the eye. Error Level Analysis [9] re-saves an image as JPEG at a known quality and measures the pixel-wise difference from the original. Regions that have been compressed differently, or that carry different noise and texture statistics, show different error levels. ELA was first used to find edited regions in photographs. Its use in this project is as an input transformation that highlights compression artifacts and inconsistencies, which may differ between machine-generated and hand-made artwork.

ELA is simple and cheap to compute, and it produces an image that a CNN can take directly. Its limits should be stated plainly: it depends on compression behaviour, so its value for AI-art detection must be checked experimentally and not assumed. This is one reason the proposed work includes an ablation study with and without ELA.

D. Explainable AI for Detection

A detector that only returns a label is hard to trust, especially when a collector or gallery must justify a decision. Explainable AI methods address this. Grad-CAM [17] uses the gradients flowing into the last convolutional layer to produce a heat map of the regions that mattered for a prediction, and LIME [18] explains single predictions by fitting a simple local model around them. Wang et al. [7] combine Grad-CAM and LIME with their transformer detector, and ArtBrain [8] adds explainability to its classification and attribution pipeline.

Explanations, however, are still treated as an add-on in most detection papers. They are rarely tied to the forensic features that the model uses, and they are not shown to the end user together with a confidence score.

E. Datasets and Generative Models

Progress depends on data. AI-ArtBench, introduced with ArtBrain [8], gathers over 185,000 images, some painted by people and some produced by generators, making it one of the largest collections available for this task, and both [5] and [6] draw on it. A wide-ranging study of synthetic-media methods in [4] covers GANs, VAEs and diffusion models across several media types, and discusses how well detectors perform, how robust they are and how they can be deployed, recommending that detection be paired with complementary technologies.

F. Summary of Gaps

Table I summarises the main studies and Table II compares them with the proposed framework. The pattern is clear: CNN studies are simple and accurate on one dataset, transformer studies are more accurate but do not use forensic features, and explainability is present in only a few of them. Very few works join forensic features, a hybrid model and explanations in one system for artwork.

TABLE I
SUMMARY OF SURVEYED STUDIES

Ref.	Year	Method	Reported result
[7]	2026	DINOv2 self-distilled transformer, Grad-CAM, LIME	99.01% accuracy; F1 94.93%
[5]	2025	LR, SVM, MLP, CNN	97.58% binary accuracy
[8]	2024	AttentionConvNeXt with XAI	≈99% generator identification
[6]	2024	CNN, VGG-19, ResNet-50	92.69% (best: CNN)

TABLE II
COMPARISON OF EXISTING SYSTEMS WITH THE PROPOSED FRAMEWORK

Feature	[7]	[5]	[8]	[6]	Proposed
AI vs. human classification	Yes	Yes	Yes	Yes	Yes
Deep learning model	Yes	Yes	Yes	Yes	Yes
Global (transformer) features	Yes	No	Part	No	Yes
Forensic ELA features	No	No	No	No	Yes
Visual explanation (XAI)	Yes	No	Yes	No	Yes

Entries reflect what each cited paper reports; "Part" means attention-based but not a full transformer.

III. PROPOSED METHODOLOGY

The proposed system is a nine-stage pipeline: data collection, image preprocessing, ELA, feature extraction, hybrid deep learning, saving of the trained model, classification, explainable AI and result presentation. Figure 1 shows the overall architecture and Figure 2 shows the ELA step in detail.

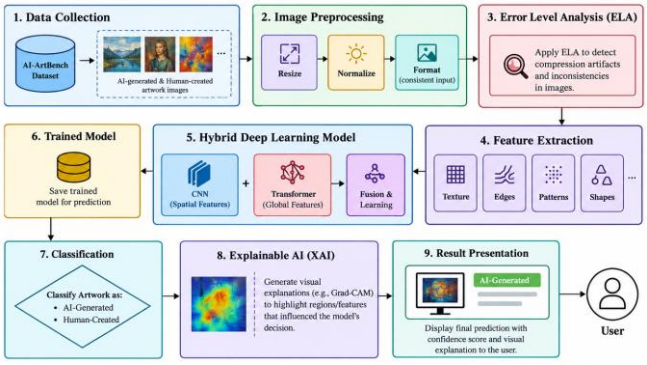

Fig. 1. Overall architecture of the proposed AI-generated artwork detection system.

A. Data Collection and Preprocessing

Training and test images will come from the AI-ArtBench dataset [8], whose human-painted and generated pictures supply the two classes needed for supervised learning. A stratified split into training, validation and test sets will be used, and images of the same artwork will not appear in more than one set. Every image is then resized to a fixed input size, its pixel values are normalised, and it is converted to one consistent format so that the later stages receive uniform input.

B. Error Level Analysis

ELA [9] re-saves each preprocessed image as a JPEG at a fixed quality Q, subtracts it from the original and scales the result, as in (1):

$$E(x, y) = s \cdot | I(x, y) - I'(x, y) | \qquad (1)$$

where I is the original image, I′ is the re-compressed image and s is a scaling factor that makes small differences visible. The map E exposes compression artifacts and inconsistencies. Figure 2 illustrates the steps. Both the ELA image and the original image will be tested as inputs, so that the contribution of ELA can be measured.

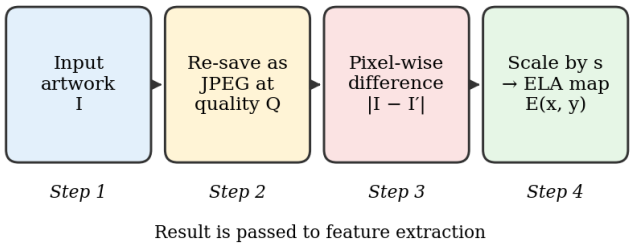

Fig. 2. Error Level Analysis process used for forensic feature extraction.

C. Feature Extraction

From the ELA output the model learns visual cues such as texture, edges, patterns and shapes. In the proposed design these features are learned by the network and are not hand-crafted, so the same stage supplies local detail to the CNN branch and image-wide structure to the transformer branch.

D. Hybrid Deep Learning Model

The hybrid model has two parallel branches followed by a fusion stage. The CNN branch captures spatial features, such as brush texture and fine local artifacts. The transformer branch [14] captures global features,

such as the relationship between distant regions of the canvas. The two feature sets are concatenated and passed to fusion layers that learn the final representation. The idea follows the finding that global and local features together help detection [7], while ELA adds a forensic signal that the earlier models did not use. After training, the best model is saved for prediction.

E. Classification and Explainable AI

A fully connected output layer classifies an artwork as AI-generated or human-created and also returns a confidence score. To explain the decision, Grad-CAM [17] is applied to the trained model and produces a heat map that marks the regions or features that pushed the prediction. LIME [18] can be added as a second check. The final screen shows the label, the confidence score and the visual explanation together, so that the user can judge the result.

F. Implementation Setup

The prototype will be built with common open-source tools. Table III lists the planned configuration.

TABLE III
SYSTEM REQUIREMENTS

Category	Configuration (planned)
Software	Python 3.10, PyTorch or TensorFlow, OpenCV, Pillow, Grad-CAM library, Windows 11 / Ubuntu
Hardware	Intel Core i7 or equivalent, 16 GB RAM, NVIDIA GPU with at least 8 GB VRAM, 512 GB storage
Dataset	AI-ArtBench (human-created and AI-generated artwork)

IV. RESULTS AND DISCUSSION

At this stage the framework has been designed and the literature reviewed; implementation and training are still to be carried out. This section therefore states what the completed system is expected to show and how it will be judged, and does not report new experimental numbers.

The first expected outcome is classification quality. The model will be measured with accuracy, precision, recall, F1-score, ROC-AUC and a confusion matrix, and compared with the published figures of the surveyed studies: 99.01% accuracy in [7], 97.58% in [5] and 92.69% in [6]. Because those works use different splits and class definitions, the comparison will also be repeated by training baseline CNN and transformer models on the same data split.

The second outcome concerns the design choices. An ablation study will train (i) a CNN alone, (ii) a transformer alone, (iii) the hybrid model on original images and (iv) the hybrid model on ELA images, so that the real benefit of ELA and of the hybrid design is measured and not assumed. A further test will train on some AI generators or art styles and test on others, to check how well the detector generalises, which is the main weakness of earlier work.

The third outcome is the usefulness of the explanations. Grad-CAM maps will be inspected on correctly and wrongly classified images to see whether the highlighted regions are consistent and meaningful, and a small group of users such as art students and artists will be asked whether the explanation helps them trust or question the result. Table IV lists the measures.

TABLE IV
EVALUATION METRICS

Metric	What it checks	Planned use
Accuracy	Share of all artworks classified correctly	Compare with [7], [5], [6]
Precision / Recall	Errors on each class	Per-class report
F1-score	Balance of precision and recall	Model selection
ROC-AUC	Separation of the two classes	Threshold study
Cross-generator test	Performance on unseen generators	Generalisation
Grad-CAM review	Consistency of highlighted regions	Trust and debugging

Compared with CNN-only detectors [5], [6], the hybrid design should capture both local and global patterns. Compared with transformer detectors [7], the proposed framework contributes an ELA-based forensic input and presents the explanation alongside the final verdict.

Four limitations must be stated. First, ELA depends on compression behaviour, so its benefit on images that have been resized, re-compressed or shared through social media is uncertain. Second, results on AI-ArtBench may not carry over to new generators, and the detector will need retraining as generators change. Third, heat maps from Grad-CAM show where the model looked, which is not always why it decided, and they should be read as aids and not as proof. Fourth, a prediction is a probability and not a legal verdict about authorship; human experts should confirm important decisions.

V. CONCLUSION

This paper reviewed the current literature on detecting AI-generated artwork and used that review to design a detection framework. The review showed that CNN classifiers are simple and accurate on a single dataset, that transformer-based detectors reach higher accuracy and sometimes give explanations, and that datasets such as AI-ArtBench make large-scale study possible, but that forensic features, hybrid modelling and explanations are seldom brought together. The proposed framework joins them: ELA for forensic features, a CNN–transformer hybrid for classification and Grad-CAM for transparent explanations. Evaluation measures have been defined, and the prototype is to be built and tested. Future work will test newer image generators, extend the dataset with more styles and sources, study robustness to resizing and re-compression, and add a user-facing web tool that returns the label, the confidence score and the explanation for any uploaded artwork.

REFERENCES

[1] I. J. Goodfellow et al., "Generative adversarial nets," in Advances in Neural Information Processing Systems (NeurIPS), vol. 27, 2014, pp. 2672–2680.

[2] J. Ho, A. Jain and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, 2020, pp. 6840–6851.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.

[4] A. I. Iliev, "Discovery of deepfakes in art," 2025.

[5] M. Li and M. Stamp, "Detecting AI-generated artwork," arXiv preprint arXiv:2504.07078, 2025.

[6] S. R. Chinta, Geethanajali, S. Kamineni, R. P. Chatragadda and K. Sujatha, "Analyzing image classification on AI-generated art vs human created art using deep learning models," in Proc. 3rd Int.

Conf. Electrical, Electronics, Information and Communication Technologies (ICEEICT), 2024, doi: 10.1109/ICEEICT61591.2024.10718485.

[7] Y. Wang, L. Li, W. Ma and Y. Zhang, "AI-generated artwork detection using self-distilled transformers with global–local feature learning and Grad-CAM interpretability," Scientific Reports, vol. 16, Art. no. 497, 2026.

[8] R. S. R. Silva, A. Lotfi, I. K. Ihianle, G. Shahtahmassebi and J. J. Bird, "ArtBrain: An explainable end-to-end toolkit for classification and attribution of AI-generated art and style," arXiv preprint arXiv:2412.01512, 2024.

[9] N. Krawetz, "A picture's worth: Digital image analysis and forensics," Black Hat Briefings, 2007.

[10] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[11] N. Tinago, S. F. Verkijika and K. E. Mamabolo, "Deepfakes in visual art: Differentiating AI-generated art from human art using convolutional neural networks (CNN)," IEEE Access, vol. 13, pp. 141484–141495, 2025, doi: 10.1109/ACCESS.2025.3596882.

[12] A. Abd-Muti Alrawahneh, S. N. H. Sheikh Abdullah, A. N. Abdul Wahab, S. K. Taylor and N. R. Nik Ab. Rahim, "A comparative study of pre-trained CNN architectures for detecting AI-generated versus human-created images," Malaysian Journal of Cybersecurity and Applications, vol. 1, no. 1, 2025, doi: 10.22452/mjca.vol1no.1.3.

[13] S. S. Baraheem and T. V. Nguyen, "AI vs. AI: Can AI detect AI-generated images?," Journal of Imaging, vol. 9, no. 10, Art. no. 199, 2023, doi: 10.3390/jimaging9100199.

[14] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.

[15] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell and S. Xie, "A ConvNet for the 2020s," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11976–11986.

[16] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," arXiv preprint arXiv:2304.07193, 2023.

[17] 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," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618–626.

[18] M. T. Ribeiro, S. Singh and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.