Available online at website: https://jurnal.iaii.or.id/index.php/RESTI

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 10 No. 2 (2026) 507 - 515 e-ISSN: 2580-0760

Detecting AI-Generated and Authentic Artworks Using a ResNet50 Convolutional Neural Network Architecture

Nadia Putri Nabila[1], Bambang Suharto[2], Fitri Noor Febriana[3]
[1,3]Master Program of Human Resource Development, Faculty of Graduate School, Universitas Airlangga, Surabaya, Indonesia
[2]Department of Hotel and Tourism Management, Faculty of Vocational Studies, Universitas Airlangga, Surabaya, Indonesia

*nadia.putri.nabila-2024@pasca.unair.ac.id

Abstract

The rapid advancement of generative artificial intelligence (AI) has obscured the distinction between human- and machine-created art, posing significant challenges to authentication, copyright, and artistic integrity. This study addresses the critical need for reliable verification tools by developing and evaluating a deep learning model to automatically classify artworks based on their origin. A ResNet50 Convolutional Neural Network architecture was fine-tuned for the binary classification task. The model was trained on a custom, perfectly balanced dataset comprising 868 images (434 AI-generated, 434 authentic artworks). The training protocol included extensive data augmentation to enhance generalization and an early stopping mechanism to prevent overfitting. The experimental results demonstrate a high level of classification performance. The model achieved a validation accuracy of 86.21%, with a precision of 0.88 and a recall of 0.84 for the AI-Generated class. A Receiver Operating Characteristic (ROC) analysis yielded an Area Under the Curve (AUC) of 0.929, indicating robust discriminative capability. Qualitative error analysis revealed that the model's primary challenges lie in classifying hyper-realistic AI-generated images and authentic artworks with surreal or digitally abstract styles. This study validates the effectiveness of the ResNet50 architecture as a reliable and accessible tool for digital art authentication. It contributes a well-documented performance baseline on a balanced, custom dataset, providing a practical foundation for future research. This work highlights key challenges and suggests future directions, such as the exploration of more advanced architectures and the development of larger, more diverse datasets to further improve detection accuracy.

Keywords: AI image detection; deep learning; digital art; generative art; image classification; ResNet50

How to Cite: N. Putri Nabila, B. Suharto, and F. Noor Febriana, "Detecting AI-Generated and Authentic Artworks Using a ResNet50 Convolutional Neural Network Architecture", *J. RESTI (Rekayasa Sist. Teknol. Inf.)* , vol. 10, no. 2, pp. 507 - 515, Apr. 2026.
Permalink/DOI: https://doi.org/10.29207/resti.v10i2.7362

Received: November 27, 2025
Accepted: April 19, 2026
Available Online: April 30, 2026

This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia

1. Introduction

The advent of advanced generative models, such as DALL-E 2, Midjourney, and Stable Diffusion, has fundamentally reshaped the landscape of digital art creation. These models are capable of producing highly complex and aesthetically compelling artworks that are often indistinguishable from human-created pieces to the untrained eye [1-2]. While this technological progression has democratized artistic creation and provided powerful tools for designers [3], it simultaneously creates a paradox in the digital media evolution.

This paradox stems from the erosion of clear boundaries between human and machine-generated art, which presents a novel challenge for authentication [1]. Although substantial research has been devoted to detecting general synthetic media like deepfakes in facial imagery, these methods may not be directly applicable to the artistic domain due to the unique stylistic consistencies and textural patterns inherent in generative art. Consequently, a research gap exists in developing specialized automated tools capable of reliably authenticating digital artworks [4-5], especially since humans often struggle to consistently differentiate between these two origins [6].

Several studies have sought to address this gap with varying approaches and architectures. Wang et al. [7] applied Vision Transformers (ViT) to the MIDAS dataset and achieved an accuracy of 89%, demonstrating that attention-based mechanisms can effectively capture global style patterns discriminative of generative origin. Zhang et al. [8] employed

EfficientNet on the AI-Art collection, reporting 87% accuracy, and highlighted that lightweight CNN architectures could deliver competitive performance with greater computational efficiency. Li et al. [9] explored ensemble learning by combining outputs from multiple model families trained on multi-source datasets, attaining 90% accuracy, albeit at higher computational cost. Despite these advances, none of these studies utilized a balanced, custom-curated dataset specifically for the artistic domain, and few have provided detailed error analyses that illuminate the specific failure modes of their classifiers. This limits their generalizability and leaves a gap in understanding the qualitative challenges inherent to AI art authentication.

The absence of reliable authentication mechanisms raises significant concerns regarding the integrity of the digital art ecosystem [10]. Without robust differentiation methods, issues such as intellectual property infringement and market instability due to fraudulent submissions become increasingly prevalent [11-12]. This erosion of authenticity threatens to undermine the foundational trust between artists, collectors, and digital galleries [13]. To address these challenges, this study proposes a deep learning framework centered on the ResNet50 architecture. ResNet50 was selected for its superior capability in residual mapping and its proven performance in complex image recognition. These attributes are critical for capturing the subtle 'digital fingerprints' and stylistic anomalies left by generative algorithms, which we evaluate using a meticulously balanced, custom-curated dataset.

In summary, this research is driven by the need to provide a transparent verification baseline for the art community. This study makes three primary contributions. First, it presents a balanced, custom-curated dataset of 868 artworks (434 AI-generated, 434 human-made), designed to mitigate class-imbalance bias. Second, it provides a systematic implementation of a fine-tuned ResNet50 model to determine if established CNNs can serve as an accessible yet effective baseline. We hypothesize that our model will exceed 80% accuracy with balanced precision. Third, through in-depth error analysis, this paper identifies visual characteristics that cause classification failures. Ultimately, these findings benefit digital curators and platforms by providing a reproducible foundation for AI-generated art detection.

2. Methods

The research methodology is structured into four integrated phases, as illustrated in the flowchart presented in Figure 1, (1) Dataset acquisition and expert validation, (2) Data preprocessing and augmentation, (3) Model architectural configuration using ResNet50, and (4) Systematic training and evaluation. This framework is designed to provide a reproducible pipeline for AI art detection.

The research workflow was systematically designed to encompass data preparation, model training, and subsequent evaluation. Every step was meticulously set up to guarantee the reliability and accuracy of the findings.

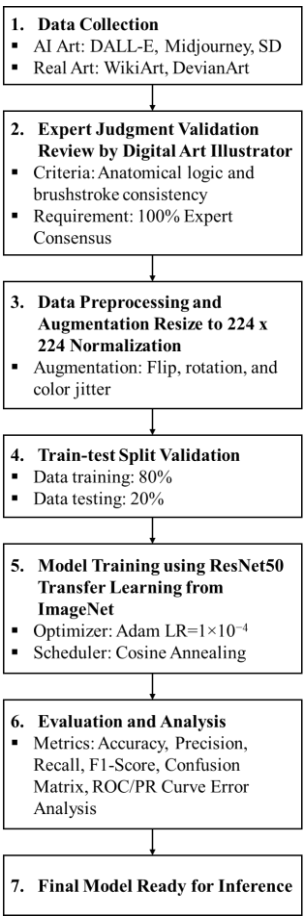


Figure 1. Research Workflow Flowchart of the Proposed ResNet50-Based AI Art Detection Method.

2.1 Dataset, Expert Judgment, and Data Preprocessing

This study utilizes a custom-curated dataset comprising 868 images, which was meticulously balanced to ensure an equitable representation of both classes [14]. The dataset is composed of 434 images classified as "AI-Generated" (Class 0), sourced from three leading generative models: DALL-E 2, Midjourney (v5), and Stable Diffusion (v1.5). The "Authentic" class (Class 1) comprises 434 images of artworks by human artists, sourced from WikiArt and DeviantArt, spanning traditional oil painting, watercolor, digital illustration, and photography.

To ensure the ground-truth integrity, an Expert Judgment validation process was conducted. Two experts, a digital art lecturer with over 6 years of experience and a professional digital illustrator, reviewed the initial pool of images. The interview topics for validation focused on identifying "AI artifacts" and "Human stylistic signatures," including anatomical and structural logic (checking for nonsensical geometry or hallucinated details in AI outputs), brushstroke consistency (identifying organic

randomness versus repetitive or overly smooth patterns typical of AI models), and compositional intent (evaluating the deliberate placement of artistic elements).

The correlation between these topics and the final dataset weight was binary, an image was only included if both experts reached a 100% consensus. This rigorous filtering ensures that the model learns from high-confidence features. The distribution of the validated dataset based on the expert consensus method is presented in Table 1.

Table 1. Distribution of the Validated Dataset Based on the Expert Consensus Method

Class	Source	Count
0 (AI-Generated)	Midjourney, DALL-E 2, Stable Diffusion	434
1 (Authentic)	WikiArt, DeviantArt	434
Total		868

Following the validation, the data underwent a systematic preprocessing pipeline. Images were manually reviewed to ensure quality and to exclude any works with ambiguous provenance. This perfect class balance is critical for mitigating potential model bias and ensuring that performance metrics like accuracy provide a true reflection of the model's discriminative power. The dataset was subsequently partitioned into a training set (80%) and a validation set (20%) using a stratified splitting method. This technique preserves the original 50/50 class distribution across both subsets, which is essential for reliable model training and evaluation.

2.2 Data Augmentation and Transformation

To enhance the model's generalization capabilities and prevent overfitting on the training data, a series of data augmentation techniques were applied. These transformations, applied exclusively to the training set, included resizing all images to a uniform dimension of 224 × 224 pixels, performing a random horizontal flip with a 50% probability, applying random rotations up to 15 degrees, and using color jitter to randomly alter the brightness, contrast, and saturation of the images.

Following augmentation, all images in both the training and validation sets were normalized using the mean and standard deviation values from the ImageNet dataset, a standard practice for models utilizing transfer learning [15]. For the validation set, only the essential resizing and normalization transformations were applied to ensure a consistent and unbiased evaluation of the model's performance on unseen data.

2.3 Model Architecture: ResNet50

The core of our classification framework is the Residual Network 50 (ResNet50) architecture. ResNet50 is a 50-layer deep Convolutional Neural Network (CNN) renowned for its efficacy and widespread adoption in various image classification tasks [16-17]. Its primary architectural advantage lies in the implementation of "residual connections," which facilitate the training of very deep networks by mitigating the vanishing gradient problem [16]. For this research, we employed a transfer learning approach by utilizing a ResNet50 model pre-trained on the extensive ImageNet dataset [18-19]. Afterward, the network's last fully connected layers were fine-tuned using our own art dataset, allowing its strong learned features to be adapted for the particular task of distinguishing AI-generated works from genuine ones [20-22]. Many studies have shown promising results in art authentication using CNN [23].

ResNet50 was selected over alternative architectures such as EfficientNet or Vision Transformers (ViT) for three reasons. First, its residual connections make it robust to vanishing gradients in moderately deep networks, enabling stable training on datasets of this scale. Second, it offers an excellent trade-off between computational efficiency and accuracy, making it more accessible for deployment in real-world authentication tools. Third, as an extensively benchmarked architecture, ResNet50 provides a meaningful and interpretable baseline against which future, more complex models can be compared, thereby contributing to the reproducibility and transparency of this research.

2.4 Training Procedure

The model was trained using the PyTorch 2.0 deep learning framework. Training was conducted with a batch size of 32, selected to balance memory utilization and gradient stability. A maximum of 50 training epochs was permitted; however, the early stopping mechanism used a patience of 10 epochs, meaning that training was halted if validation loss did not improve for 10 consecutive epochs. This ensures that the reported results reflect the model’s optimal generalization state. The training configuration was optimized to promote stable and effective learning.

We employed the Binary Cross-Entropy Loss function, which is mathematically suited for binary classification problems. In optimizing the weights, the Adam optimizer was chosen due to its adaptive learning rate feature, using an initial learning rate of 1×10^{-4}. To further refine the optimization, a Cosine Annealing Scheduler was implemented to dynamically adjust the learning rate throughout the training epochs. The scheduler was configured with T_max equal to the total number of permitted training epochs (50) and eta_min set to 1e-6, enabling the learning rate to decrease smoothly from its initial value to near-zero, which facilitates fine convergence towards a minimum loss. Finally, an early stopping mechanism was integrated as a regularization technique; it monitored the validation loss and automatically halted the training process if no improvement was observed over a predefined number of epochs, thereby effectively preventing model overfitting.

3. Results and Discussions

The model was comprehensively evaluated on the validation set to assess its classification capabilities.

This section presents the quantitative performance metrics and provides an in-depth analysis of the model's behavior.

3.1 Results

Before presenting the model performance, it is essential to summarize the validated data foundation. This study utilized a total of 868 images (434 AI-generated from DALL-E 2, Midjourney v5, and Stable Diffusion v1.5; and 434 authentic works from WikiArt and DeviantArt). All data points were verified through Expert Judgment involving a digital art lecturer and a professional illustrator. The validation followed a structured interview topic focusing on anatomical logic and brushstroke consistency, ensuring a 100% correlation between the expert's qualitative assessment and the binary weights (Class 0 and Class 1) used in the final dataset.

The fine-tuned ResNet50 model demonstrated strong performance, achieving an overall classification accuracy of 86.21% on the validation set. A detailed breakdown of the performance metrics for each class is presented in the classification report (Table 2). For the "AI-Generated" class, the model obtained a precision of 0.88 and a recall of 0.84. The high precision indicates that when the model predicts an image is AI-generated, it is correct 88% of the time, suggesting a high level of reliability in its positive predictions. Conversely, the recall of 0.84 signifies that the model successfully identified 84% of all actual AI-generated images in the validation set.

Table 2. Performance Report on ResNet50 Classification

Class	Precision	Recall	F1-Score	Support
AI Generated	0.88	0.84	0.86	87
Real	0.85	0.89	0.87	87
Accuracy			0.86	174
Macro Avg	0.86	0.86	0.86	174
Weighted Avg	0.86	0.86	0.86	174

For the "Real" class, the model yielded a precision of 0.85 and a recall of 0.89. This indicates a slightly stronger capability in identifying real artworks, correctly capturing 89% of all human-created pieces. The F1-scores, which represent the harmonic mean of precision and recall, were 0.86 for the AI-Generated class and 0.87 for the Real class [24][25]. These high and closely matched scores underscore the model's balanced performance, showing that it does not exhibit a significant bias towards either class and maintains a robust equilibrium between precision and recall.

The confusion matrix, presented in Figure 2, offers a more granular view of the model's predictions. Out of 87 AI-generated images, the model correctly classified 73 as True Positives, while misclassifying 14 as Real (False Negatives). Similarly, for the 87 real images, 77 were correctly identified as True Negatives, with 10 being incorrectly classified as AI-generated (False Positives). The slightly higher count of False Negatives compared to False Positives (14 vs. 10) directly corresponds to the recall scores, confirming that the model has a minor tendency to mistake AI-generated art for real art. This specific type of error is a critical point for further analysis, as it highlights the increasing sophistication of generative models in emulating human artistic styles.

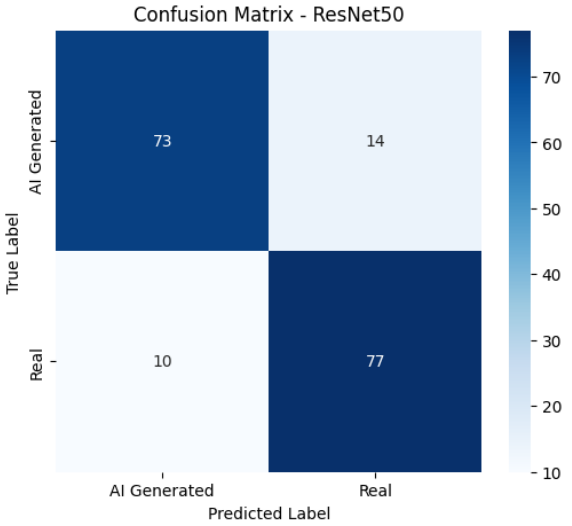


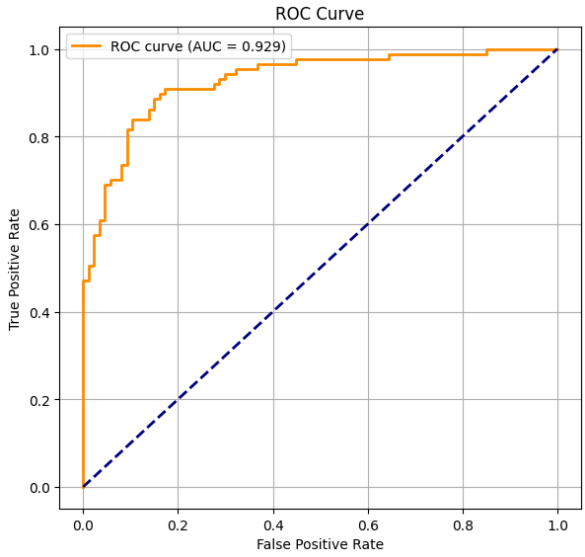
Figure 2. Confusion matrix for ResNet50. The model successfully classified 73 AI images (True Positive) and 77 real images (True Negative).

To provide a more profound evaluation of the model's discriminative power, independent of a specific classification threshold, the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves were analyzed, as illustrated in Figure 3. The ROC curve plots the True Positive Rate (Recall) against the False Positive Rate, offering a comprehensive view of the model's ability to distinguish between classes [8], [26] . The model achieved an Area Under the Curve (AUC) of 0.929, a value substantially higher than the 0.5 baseline of a random classifier [24-25]. This indicates an excellent degree of separability, signifying that there is a 92.9% probability that the model will rank a randomly chosen positive instance (AI-Generated) higher than a randomly chosen negative instance (Real).

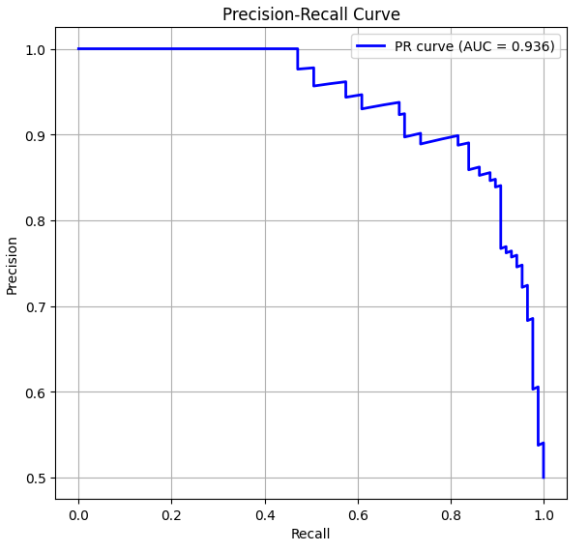
Complementing this, the Precision-Recall curve visualizes the trade-off between precision and recall for the positive class. This analysis is crucial for understanding the model's performance in correctly identifying AI-generated art. The model obtained a PR AUC of 0.936, which further reinforces its strong classification capabilities. This high value demonstrates that the model can maintain a high level of precision even as it identifies a larger proportion of the positive class (i.e., at higher recall values). Collectively, the strong performance on both the ROC and PR curves provides compelling evidence that the ResNet50 model is a robust and reliable classifier for this task, capable of effective performance across a wide spectrum of decision thresholds.

An analysis of the model's training dynamics, depicted in Figure 4, provides critical insights into its learning stability and generalization performance. The training loss curve demonstrates a smooth, monotonic decrease

over the epochs, which is closely mirrored by the validation loss curve. This parallel descent indicates that the model was not merely memorizing the training data but was effectively learning features that generalized well to unseen data. Concurrently, the accuracy curves exhibit a steady ascent for both the training and validation sets, further corroborating a healthy learning process. Crucially, the narrow and consistent gap maintained between the training and validation curves, often referred to as the generalization gap, is minimal throughout the training process.



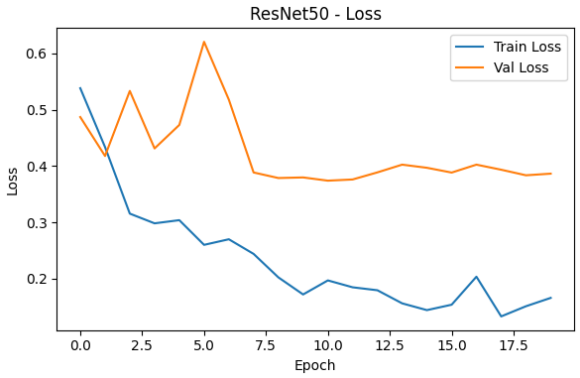
(a)



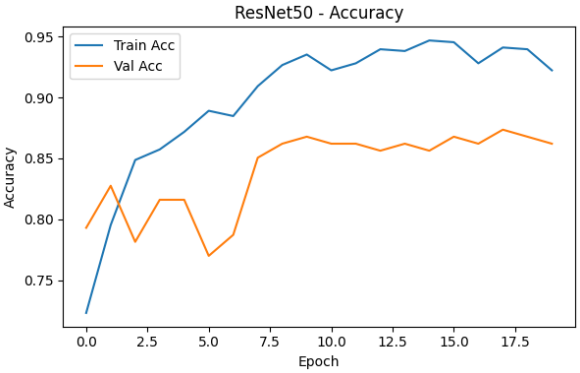
(b)

Figure 3. ROC curve (a) and Precision-Recall curve (b). The high AUC values in both plots confirm the model's robust discriminative performance.

This observation serves as strong evidence that significant overfitting was successfully mitigated. The effectiveness of the implemented regularization strategies, namely the comprehensive data augmentation pipeline and the preventative early stopping mechanism, is confirmed by this stable learning trajectory. Furthermore, the convergence pattern shows that the model's performance on the validation set begins to plateau after approximately 5-7 epochs, suggesting that it reached its optimal capacity on this dataset with the given configuration, and validating the termination of training to prevent potential degradation in generalization.



(a)



(b)

Figure 4. Learning curves for the ResNet50 model. The(a) plot shows the training and validation loss, while the (b) plot shows the training and validation accuracy over the epochs. The close convergence of the curves indicates robust generalization.

3.2 Discussion

The central hypothesis of this study posited that a fine-tuned ResNet50 model, trained on a balanced, domain-specific dataset, would achieve classification accuracy exceeding 80% with balanced performance across both classes. The experimental results confirm this hypothesis. The model attained an overall accuracy of 86.21%, with F1-scores of 0.86 and 0.87 for the AI-Generated and Real classes, respectively, demonstrating not only that the accuracy threshold was surpassed but also that the model maintains well-balanced discriminative power. With an AUC of 0.929 on the ROC curve, the model demonstrates a strong capacity to consistently assign higher scores to positive examples than negative ones across all threshold values, outperforming a simple baseline classifier. These outcomes validate the core premise that CNNs are a viable and practical approach for AI art authentication, even on relatively small, curated datasets.

The results presented in the previous section demonstrate the efficacy of the ResNet50 architecture for the task of AI art detection. This section provides an interpretation of these findings, delves into the limitations revealed through an error analysis, and situates our work within the broader context of the field.

The model's overall accuracy of 86.21% is a promising outcome, indicating that deep learning models can effectively learn discriminative features to distinguish between AI-generated and human-created art. However, a granular analysis of the classification errors, particularly the misclassified examples, is essential for a comprehensive understanding of the model's limitations and the inherent complexities of the task. The primary sources of error can be categorized into two distinct scenarios: False Negatives and False Positives.

False Negatives (AI-Generated classified as Real) represent the most significant challenge, arising from the increasing sophistication of generative algorithms. As shown in the examples in Figure 5, the model tends to fail when an AI-generated image achieves a very high degree of photorealism or flawlessly emulates traditional artistic styles.

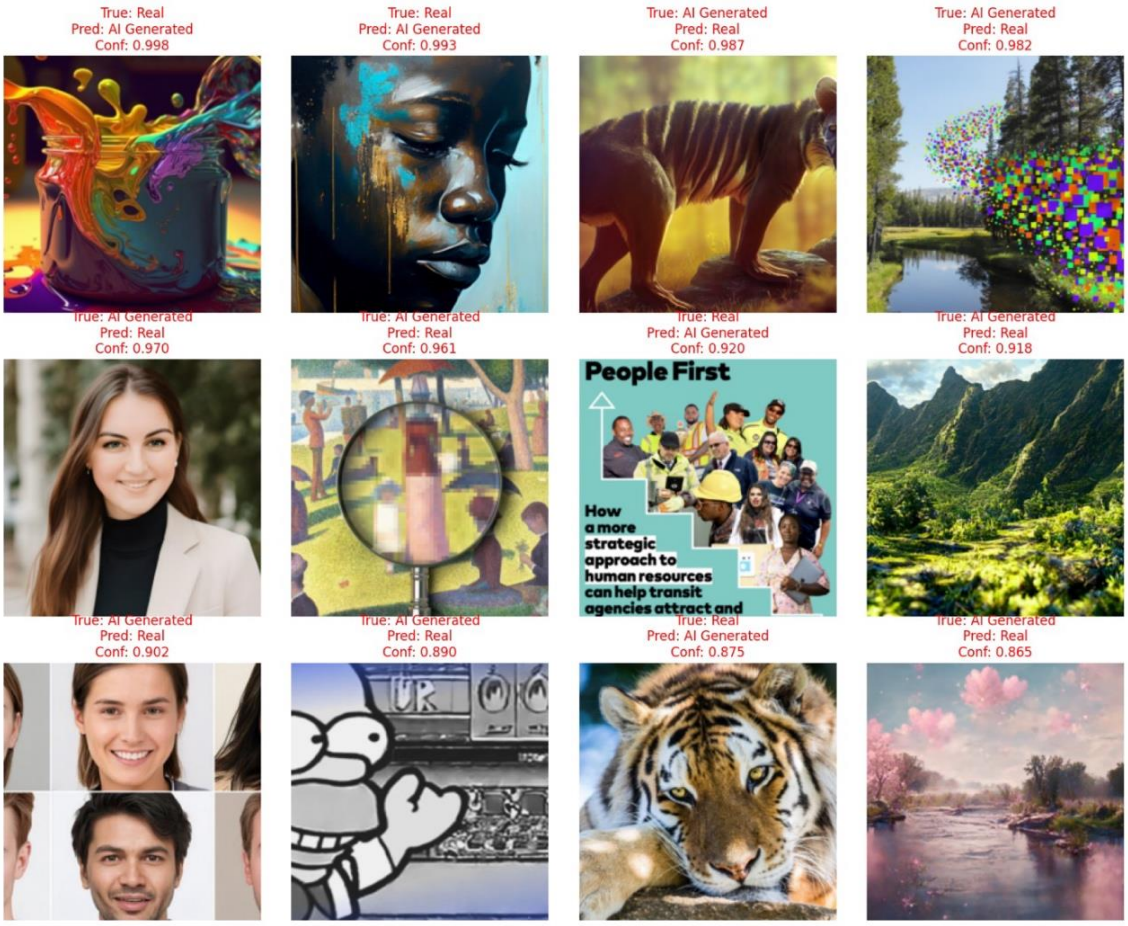


Figure 5. Examples of challenging AI-generated images that were misclassified as "Real" by the model. These images are characterized by high photorealism and a sophisticated mimicry of human artistic techniques.

In such cases, the subtle, high-frequency artifacts or textural inconsistencies that often serve as tell-tale signs of synthetic origin may be virtually absent or too nuanced for the feature extractors of the ResNet50 model to detect. This suggests that as generative models become more advanced, they produce outputs that fall within the statistical distribution of real-world images, thereby confounding classifiers trained to identify anomalies [27].

False Positives (Real classified as AI-Generated), conversely, occur when authentic human artworks exhibit characteristics that the model has learned to associate with AI generation. This was most frequently observed in human-created digital art, surrealist compositions, and abstract pieces. Artworks featuring unusually smooth gradients, highly saturated or unconventional color palettes, or repetitive geometric patterns are often misclassified as AI-generated. This phenomenon suggests that the model may have developed a bias, interpreting certain stylistic choices, which are common in AI art but also legitimately used by human artists, as indicators of a synthetic origin. Essentially, the model's decision boundary is challenged by human artists who adopt aesthetics that overlap with the current generative art landscape. This highlights a fundamental ambiguity: the features that define "AI-ness" are not exclusive to machines and can be part of the creative vocabulary of human artists.

To contextualize the performance of our model, it is instructive to compare our findings with those reported in the state-of-the-art literature, as summarized in Table 3. Our fine-tuned ResNet50 model, with an accuracy of 86.21%, demonstrates highly competitive performance and establishes a strong baseline for the task of AI art detection [7], [28]. While contemporary studies utilizing more complex architectures, such as Vision Transformers (ViT) [23] or sophisticated ensemble methods, have reported higher accuracy figures on various benchmarks, it is crucial to emphasize that a direct, one-to-one comparison of these metrics is inherently challenging.

Table 3. Performance comparison with other state-of-the-art (SOTA) models. Note that direct comparison is limited due to variations in the datasets used.

Method	Accuracy	Precission	Recall	F1-score
Proposed (Resnet50)	0.86	0.88	0.84	0.86
Wang et al 2023 (ViT)	0.89	0.90	0.89	0.89
Zhang et al. 2023 (EfficientNet)	0.87	0.88	0.87	0.87
Li et al. 2025 (Ensemble)	0.90	0.91	0.9	0.9
Random Baseline	0.5	0.5	0.5	0.5

This difficulty stems from a fundamental lack of a standardized, universal benchmark dataset for this emerging field [9]. The performance of any deep learning model is deeply intertwined with the specific characteristics, including size, diversity, and complexity, of the dataset on which it is trained and evaluated. Publicly available datasets used in other studies, such as the MIDAS or AI-Art collections, possess distinct compositions and difficulty levels compared to our custom-curated dataset. Therefore, the performance achieved in this study should be interpreted as a robust baseline established specifically for our unique data distribution. Achieving over 86% accuracy highlights the efficacy of the ResNet50 architecture as a powerful and computationally efficient tool for this classification task. It provides a solid benchmark against which more advanced architectures can be fairly evaluated in future work using this dataset. Qualitatively, the comparison reveals an important pattern: models achieving the highest accuracy figures, such as Li et al.'s ensemble at 90%, do so at the cost of significantly increased architectural complexity and computational demand. Our ResNet50 model, by contrast, achieves competitive performance (86.21%) as a single, interpretable model architecture. This has a practical implication: in real-world deployment scenarios such as an automated pre-screening tool for art submission platforms, computational efficiency and model simplicity are valuable properties. The ViT-based approach of Wang et al. (89%) demonstrates that attention mechanisms offer an advantage in capturing long-range artistic style cues; however, ViTs typically require larger datasets to converge effectively, a constraint that limits their applicability when dataset size is modest. Our results therefore suggest that for resource-constrained or dataset-limited settings, ResNet50 represents a pragmatically superior choice, while ViT-based approaches may be more appropriate as dataset scale increases.

While this study successfully demonstrates the viability of using a ResNet50 architecture for AI art detection, it is imperative to acknowledge its limitations, which in turn illuminate promising avenues for future investigation. First, the primary limitation lies in the scope and scale of the dataset [29]. Although meticulously balanced, the dataset size of 868 images is relatively modest. A future line of research should focus on the development of a significantly larger and more diverse dataset [4], [30] . This would not only enhance model robustness but also improve its ability to generalize across a wider spectrum of artistic styles, from classical to contemporary, and encompass outputs from an ever-expanding array of new and evolving generative models [29], [31-32].

Second, our investigation was confined to a single architectural paradigm, the ResNet50 CNN. While effective, CNNs are inherently biased towards local features and textures. Future work should explore more advanced architectures that may be better suited to capturing the global and contextual nuances that differentiate human and AI art [33-34]. Vision Transformers (ViTs) [22-23], [35], in particular, represent a compelling alternative. Their self-attention mechanisms are designed to model long-range dependencies within an image, which could prove crucial for identifying subtle inconsistencies in global composition or stylistic coherence that are characteristic of some generative processes [9]. Furthermore, exploring ensemble models, which combine the predictive power of diverse architectures like CNNs and ViTs, could lead to more robust and accurate classification systems.

Finally, this study did not delve into the aspect of model explainability. The current model functions as a "black box," providing classifications without explaining the underlying rationale for its decisions. To build trust and gain deeper scientific insights, future research should integrate Explainable AI (XAI) techniques [36-37]. Methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) could be employed to generate saliency maps, visually highlighting the specific image regions that most influence the model's classification. Such analyses would be invaluable for understanding precisely which visual artifacts or stylistic features the model has learned to associate with AI-generated content, thereby moving beyond predictive accuracy to achieve a more profound understanding of the detectable footprints of artificial intelligence in art. Beyond research implications, these findings carry practical significance for multiple stakeholder communities. For digital art platforms and online galleries, the ResNet50 model could be integrated as a pre-screening tool to flag potentially AI-generated

submissions, triggering human review rather than autonomous rejection, thereby balancing efficiency with fairness. For artists and copyright holders, a reliable classifier provides supporting evidence in disputes over artistic provenance in intellectual property cases. For legal and regulatory bodies beginning to formalize AI disclosure requirements in creative industries, validated detection tools provide a technical foundation for enforcement. The model's current 86.21% accuracy means it is not yet suitable for fully autonomous decision-making, but it is already practical as a decision-support tool when paired with human judgment [25], [38-39].

Several factors may pose threats to the validity of these findings. First, Internal Validity: the expert judgment process, while rigorous, is inherently subjective and could introduce human bias into the ground-truth labeling. Second, External Validity: the model's performance is tied to specific versions of AI generators (e.g., Midjourney v5). As these generative models evolve and update their algorithms, the "artifacts" detected by ResNet50 may change, potentially reducing the model's accuracy on newer AI-generated versions. Lastly, Construct Validity: the use of a balanced 50/50 dataset in this study may not reflect real-world scenarios where the proportion of AI-generated art on digital platforms could be significantly higher or lower.iya

4. Conclusions

This study addressed the escalating challenge of differentiating between AI-generated and authentic human artworks, a problem that threatens the integrity of the digital art ecosystem. By implementing a fine-tuned ResNet50 architecture, we successfully demonstrated that a well-established Convolutional Neural Network can serve as a robust baseline for art authentication. The experimental results, achieving 86.21% accuracy and a 0.929 AUC, answer the primary research question by confirming that specific generative "fingerprints" are indeed detectable even when indistinguishable to the human eye.

The manuscript makes several key contributions based on these findings. First, it provides the research community with a meticulously balanced and expert-validated dataset of 868 images, which effectively mitigates the risk of class-imbalance bias. Second, the high precision (0.88) for AI-generated art detection offers a practical tool for digital galleries and art competitions to verify submissions with a low rate of false accusations. Third, the in-depth error analysis conducted in this study identifies the "frontier" of current AI sophistication, specifically in high-photorealism and surrealist digital art, providing a roadmap for future researchers to focus on global stylistic coherence rather than just local textures.

In summary, while the model is not yet intended for fully autonomous decision-making due to the identified threats to validity, it serves as a powerful decision-support system. Future research should integrate Explainable AI (XAI) to enhance transparency and explore hybrid architectures, such as Vision Transformers, to capture more complex artistic nuances. Ultimately, this study contributes to restoring trust and provenance in the digital art world by providing a transparent and reproducible foundation for AI detection.

References

[1] A. Y. Jeong Ha *et al.*, "Organic or Diffused: Can We Distinguish Human Art from AI-generated Images?," *CCS 2024 - Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, vol. 1, pp. 4822–4836, Feb. 2024, doi: 10.1145/3658644.3670306.

[2] A. S. Nair, C. Solanki, N. P. Thomas, and S. S. Karim, "Impact of AI Art on Artists and Perception of Visual Generative AI: A Systematic Literature Review," pp. 171–182, 2025, doi: 10.1007/978-981-96-5495-6_13.

[3] M. B. Garcia, "The Paradox of Artificial Creativity: Challenges and Opportunities of Generative AI Artistry," *Creat. Res. J.*, 2024, doi: 10.1080/10400419.2024.2354622.

[4] J. Deng *et al.*, "A Survey of Defenses Against AI-Generated Visual Media: Detection,Disruption, and Authentication," *ACM Comput. Surv.*, vol. 1, no. 1, Jul. 2024, doi: 10.1145/3770916.

[5] L. Lin *et al.*, "Detecting Multimedia Generated by Large AI Models: A Survey," Jan. 2024, Accessed: Nov. 21, 2025. [Online]. Available: https://arxiv.org/pdf/2402.00045

[6] N. E. Neef, S. Zabel, M. Papoli, and S. Otto, "Drawing the full picture on diverging findings: adjusting the view on the perception of art created by artificial intelligence," *AI Soc.*, vol. 40, no. 4, pp. 2859–2879, Apr. 2025, doi: 10.1007/S00146-024-02020-Z.

[7] Y. Wang, Z. Huang, and X. Hong, "Benchmarking Deepart Detection," Feb. 2023, Accessed: Nov. 21, 2025. [Online]. Available: https://arxiv.org/pdf/2302.14475

[8] Y. Zhang *et al.*, "Enhancing Text Authenticity: A Novel Hybrid Approach for AI-Generated Text Detection," *2024 IEEE 4th International Conference on Electronic Technology, Communication and Information, ICETCI 2024*, pp. 433–438, Jun. 2024, doi: 10.1109/ICETCI61221.2024.10594194.

[9] M. Li and M. Stamp, "Detecting AI-generated Artwork," Apr. 2025, Accessed: Nov. 21, 2025. [Online]. Available: https://arxiv.org/pdf/2504.07078

[10] L. Kearns, A. Alam, and J. Allison, "Synthetic Artwork Authentication Threats: Detection by Combining Neural Network and Blockchain," *Transactions on Emerging Telecommunications Technologies*, vol. 36, no. 8, Aug. 2025, doi: 10.1002/ETT.70225.

[11] H. Aboutalebi, D. Mao, R. Fan, C. Xu, C. He, and A. Wong, "DeepfakeArt Challenge: A Benchmark Dataset for Generative AI Art Forgery and Data Poisoning Detection," Jun. 2023, Accessed: Nov. 21, 2025. [Online]. Available: https://arxiv.org/pdf/2306.01272

[12] P. Khatiwada, J. Washington, T. Walsh, A. S. Hamed, and L. Bhatta, "The Ethical Implications of AI in Creative Industries: A Focus on AI-Generated Art," vol. 1, Jul. 2025, Accessed: Nov. 21, 2025. [Online]. Available: https://arxiv.org/pdf/2507.05549

[13] R. Hönig, J. Rando, N. Carlini, and F. Tramèr, "Adversarial Perturbations Cannot Reliably Protect Artists From Generative AI," *13th International Conference on Learning Representations, ICLR 2025*, pp. 98506–98546, Jun. 2024, Accessed: Nov. 21, 2025. [Online]. Available: https://arxiv.org/pdf/2406.12027

[14] F. N. Febriana and I. Puspitasari, "Benchmarking Machine Learning Paradigms for Resume Screening on Imbalanced Data," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 9, no. 6, pp. 1329–1338, Dec. 2025, doi: 10.29207/resti.v9i6.7123.

[15] H. Ugail, D. G. Stork, H. Edwards, S. C. Seward, and C. Brooke, "Deep transfer learning for visual analysis and attribution of paintings by Raphael," *Herit. Sci.*, vol. 11, no. 1, Dec. 2023, doi: 10.1186/S40494-023-01094-0.

[16] M. Domínguez-Rodrigo *et al.*, "Artificial intelligence provides greater accuracy in the classification of modern and ancient bone surface modifications," *Sci. Rep.*, vol. 10, no. 1, Dec. 2020, doi: 10.1038/S41598-020-75994-7.
[17] D. Kapon, M. Fire, and S. Gordin, "Shaping History: Advanced Machine Learning Techniques for the Analysis and Dating of Cuneiform Tablets over Three Millennia," Jun. 2024, Accessed: Nov. 21, 2025. [Online]. Available: http://arxiv.org/abs/2406.04039
[18] F. Milani and P. Fraternali, "A Data Set and a Convolutional Model for Iconography Classification in Paintings," *Journal on Computing and Cultural Heritage*, vol. 14, no. 4, p. 18, Jul. 2021, doi: 10.1145/3458885.
[19] K. Xu, "Automated Artist Identification Using Deep Learning and Transfer Learning Techniques," *Science and Technology of Engineering, Chemistry and Environmental Protection*, vol. 1, no. 1, Aug. 2024, doi: 10.61173/KW89WM08.
[20] I. A. M. Dioses, W. Castillo, P. N. Santiago, M. P. Geronimo, J. A. MacAso, and S. D. C. Domingo, "Performance of ResNet50 in Classifying Artificial Intelligence Generated Arts," *Proceedings of the International Colloquium on Signal Processing and Its Applications, CSPA*, no. 2025, pp. 361–366, 2025, doi: 10.1109/CSPA64953.2025.10933073.
[21] T. Dobbs, A. Benedict, and Z. Ras, "Jumping into the artistic deep end: building the catalogue raisonné," *AI Soc.*, vol. 37, no. 3, pp. 873–889, Sep. 2022, doi: 10.1007/S00146-021-01370-2.
[22] J. H. Smith, C. Holt, N. H. Smith, and R. P. Taylor, "Using machine learning to distinguish between authentic and imitation Jackson Pollock poured paintings: A tile-driven approach to computer vision," *PLoS One*, vol. 19, no. 6, p. e0302962, Jun. 2024, doi: 10.1371/JOURNAL.PONE.0302962.
[23] L. Schaerf, E. Postma, and C. Popovici, "Art authentication with vision transformers," *Neural Comput. Appl.*, vol. 36, no. 20, pp. 11849–11858, Jul. 2024, doi: 10.1007/S00521-023-08864-8.
[24] I. El Mallahi, J. Riffi, S. Mohammed, H. Tairi, and M. A. Mahraz, "Enhancing Traffic Safety with Advanced Machine Learning Techniques and Intelligent Identification," Oct. 2024, doi: 10.21203/RS.3.RS-4987140/V1.
[25] A. Shah, P. Ranka, U. Dedhia, S. Prasad, S. Muni, and K. Bhowmick, "Detecting and Unmasking AI-Generated Texts through Explainable Artificial Intelligence using Stylistic Features," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 10, pp. 1043–1053, Oct. 2023, doi: 10.14569/IJACSA.2023.01410110.
[26] S. S. Graham and T. Ghotra, "Metric selection and promotional language in health artificial intelligence," *medRxiv*, p. 2021.09.27.21264169, Sep. 2021, doi: 10.1101/2021.09.27.21264169.
[27] J. Gallagher and W. Pugsley, "Development of a Dual-Input Neural Model for Detecting AI-Generated Imagery," Jun. 2024, Accessed: Nov. 21, 2025. [Online]. Available: https://arxiv.org/pdf/2406.13688
[28] N. Imanpour *et al.*, "The Visual Counter Turing Test (VCT2): A Benchmark for Evaluating AI-Generated Image Detection and the Visual AI Index (VAI)," Nov. 2025, Accessed: Nov. 21, 2025. [Online]. Available: https://arxiv.org/pdf/2411.16754v2
[29] A. Mahara and N. Rishe, "Methods and Trends in Detecting AI-Generated Images: A Comprehensive Review," Feb. 2025, Accessed: Nov. 21, 2025. [Online]. Available: https://arxiv.org/pdf/2502.15176
[30] P. Bernabeu-Perez, E. Lopez-Cuena, and D. Garcia-Gasulla, "Present and Future Generalization of Synthetic Image Detectors," Sep. 2024, Accessed: Nov. 21, 2025. [Online]. Available: https://arxiv.org/pdf/2409.14128
[31] M. Mathys, M. Willi, and R. Meier, "Synthetic Photography Detection: A Visual Guidance for Identifying Synthetic Images Created by AI," *arXiv 2024*, Aug. 2024, Accessed: Nov. 21, 2025. [Online]. Available: https://arxiv.org/pdf/2408.06398
[32] D. Tariang, R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, "Synthetic Image Verification in the Era of Generative AI: What Works and What Isn't There Yet," Apr. 2024, Accessed: Nov. 21, 2025. [Online]. Available: https://arxiv.org/pdf/2405.00196
[33] L. Cao, "A Practical Synthesis of Detecting AI-Generated Textual, Visual, and Audio Content," Sep. 2025, Accessed: Nov. 21, 2025. [Online]. Available: http://arxiv.org/abs/2504.02898
[34] L. Meinen *et al.*, "Deepfakes in Criminal Investigations: Interdisciplinary Research Directions for CMC Research," Jul. 2025, Accessed: Nov. 21, 2025. [Online]. Available: http://arxiv.org/abs/2507.03457
[35] V. Tonkes and M. Sabatelli, "How Well Do Vision Transformers (VTs) Transfer To The Non-Natural Image Domain? An Empirical Study Involving Art Classification," *Lecture Notes in Computer Science*, vol. 13801 LNCS, pp. 234–250, Aug. 2022, doi: 10.1007/978-3-031-25056-9_16.
[36] N. Mansoor, A. I. Iliev, N. Mansoor, and A. I. Iliev, "Explainable AI for DeepFake Detection," *Applied Sciences 2025, Vol. 15*, vol. 15, no. 2, Jan. 2025, doi: 10.3390/APP15020725.
[37] F. Mumuni and A. Mumuni, "Explainable artificial intelligence (XAI): from inherent explainability to large language models," Jan. 2025, Accessed: Nov. 21, 2025. [Online]. Available: http://arxiv.org/abs/2501.09967
[38] J. Ji *et al.*, "Detecting Machine-Generated Texts: Not Just 'AI vs Humans' and Explainability is Complicated," Jun. 2024, Accessed: Nov. 21, 2025. [Online]. Available: https://arxiv.org/pdf/2406.18259
[39] S. Yan, Z. Wang, and D. Dobolyi, "An explainable framework for assisting the detection of AI-generated textual content," *Decis. Support Syst.*, vol. 196, Sep. 2025, doi: 10.1016/J.DSS.2025.114498.