

Paper Type: Original Article

Image Classification and Explainable Identification of AI-Generated Images

Hamed Elsimary^{1,*}  **and Mohamed Refaat Abdellah²** ¹ Department of Computer Science, Cairo Higher Institute for Engineering Computer Science and Management, Cairo, Egypt; h.elsimary@chi.edu.eg.² Department of Computer Science, College of Information Technology, Misr University for Science and Technology, Cairo, Egypt; mohamed.refaat@must.edu.eg.**Received:** 31 Jan 2025**Revised:** 13 May 2025**Accepted:** 20 Nov 2025**Published:** 24 Nov 2025**Abstract**

The unprecedented advancements in Artificial Intelligence (AI) over the past few years have significantly enhanced the ability of generative models to produce synthetic images that are virtually indistinguishable from real-world photographs. As the boundary between authentic and AI-generated images becomes increasingly blurred, the critical necessity for reliable methods of image authentication and data verification has emerged. In response to this growing challenge, the present study proposes a novel approach to enhancing the detection of AI-generated images through the application of computer vision techniques. The research introduces the creation of a synthetic dataset, named Artifact, which mirrors the ten classes found within the widely utilized Artifact dataset. This dataset was generated using a Latent Diffusion Model (LDM), capable of producing highly complex and photorealistic visual features, such as realistic reflections and depth of field effects. The binary classification task is then formulated, challenging the system to distinguish between real and synthetic imagery. To address this task, a Convolutional Neural Network (CNN) architecture was designed, trained, and rigorously optimized across 36 different configurations through hyperparameter tuning. The resulting model achieved an impressive classification accuracy of 94.30%, demonstrating its ability to discern subtle imperfections and anomalies often overlooked by human observers. Furthermore, the study integrates Explainable AI (XAI) techniques, employing Gradient Class Activation Mapping (Grad-CAM) to visualize the internal decision-making processes of the CNN model. These visual explanations reveal that, contrary to intuitive assumptions, the CNN does not primarily focus on the main objects within the images. Instead, it detects small background inconsistencies and imperfections that serve as key indicators of synthetic origin.

Keywords: AI-generated Images; Image Classification; CNN.

1 | Introduction

In recent years, the field of synthetic image generation through Artificial Intelligence (AI) has witnessed extraordinary progress. The rapid evolution of this technology has led to a crucial and urgent necessity: the ability to detect and differentiate between authentic photographic images and those created by AI. Traditionally, artificially generated images often contained glaring defects, making them easy to distinguish from real photographs. However, advancements in generative models have now reached a point where AI

**Corresponding Author:** h.elsimary@chi.edu.eg<https://doi.org/10.61356/j.scin.2025.2615>

Licensee **SciNexuses**. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0>).

can produce photorealistic images in a matter of seconds, with quality levels high enough to deceive even trained human observers.

2 | Background and Related Work

The ability to differentiate between genuine photographic content and images generated by artificial intelligence has become increasingly vital across various sectors [1]. Ensuring the authenticity and integrity of visual data is essential for trust, legal accountability, cybersecurity, and accurate journalism [2]. Synthetic imagery, if left undetected, could be weaponized to deceive the public, fabricate evidence, or facilitate unauthorized digital access [3].

While earlier AI-generated images often contained obvious flaws, recent models particularly Latent Diffusion Models have drastically improved image quality, making detection much harder [4]. These models operate by learning the reverse process of Gaussian noise diffusion, conditioned by text prompts, to generate detailed and high-fidelity images [5]. The societal risks are profound: generative models could create realistic fake news, false criminal evidence, or synthetic identities capable of bypassing security systems. Recent studies have shown that AI-generated signatures and faces can compromise authentication systems with alarming ease [6].

Although detection methods are still in their infancy, some promising approaches have emerged. Techniques based on analyzing digital "fingerprints," frequency domain artifacts, optical flow inconsistencies, and chrominance disparities have shown potential. However, the rapid advancement of generative models means that detection systems must continually evolve to keep pace [7].

The detection of AI-generated images has become an increasingly active area of research due to the rapid advancement of generative models such as Generative Adversarial Networks (GANs) and diffusion models. Early works focused on detecting artifacts and visual inconsistencies introduced by (GANs), such as unnatural textures, distorted facial features, or irregular lighting. Techniques like frequency domain analysis, Benford's Law, and patch-based inconsistencies were commonly used to highlight these discrepancies [8].

With the introduction of more advanced models such as StyleGAN and Stable Diffusion, researchers began exploring deep learning-based detection methods. CNN architectures were often employed to extract subtle features beyond human perception, while ensemble learning and transfer learning further improved detection accuracy. Some studies also incorporated Explainable AI (XAI) tools, such as Grad-CAM, to provide interpretability by highlighting which parts of an image influenced the model's prediction. Other works extended detection into multimedia contexts, including videos and deepfake audio, emphasizing the broader risks of synthetic media. Despite progress, challenges remain. As generative models become increasingly photorealistic and adaptable, the boundary between real and synthetic continues to blur demanding more robust, adaptable, and interpretable detection systems.

3 | Methodology

In this study, a comprehensive methodology was adopted to detect AI-generated synthetic images by combining advanced data generation, deep learning classification, and explainable AI techniques.

First, synthetic data was generated using a Latent Diffusion Model (LDM), specifically the publicly available Stable Diffusion v1.4 model. This model was employed to create 60,000 synthetic images corresponding to the ten classes of the Artifact dataset, using carefully crafted text prompts and diverse prompt modifiers to ensure high variability and realism. These synthetic images were resized to 32×32 pixels to match the dimensions of the real Artifact images [4].

For the classification task, a Convolutional Neural Network (CNN) architecture was designed and optimized. The CNN was trained to perform binary classification, distinguishing between real and AI-generated images. Extensive hyperparameter tuning was conducted, involving the evaluation of different numbers of

convolutional filters, layers, and dense units to achieve the highest possible accuracy while minimizing loss and classification errors [1].

To interpret the CNN's decision-making process and enhance transparency, the study integrated Explainable AI (XAI) techniques. Specifically, Gradient Class Activation Mapping (Grad-CAM) was applied to visualize the regions of the images that most influenced the model's predictions. This provided critical insights into how the CNN differentiated real from synthetic images, revealing that the model often focused on minor background imperfections rather than the primary objects themselves [7].

All experiments were conducted using the TensorFlow deep learning framework, and computational resources included an Nvidia Apple M1 GPU to handle the training and data generation tasks efficiently.

4 | Dataset Description

In this study, we employed the (Artifact) dataset, a curated image dataset designed to support research in detecting synthetic imagery generated by artificial intelligence. The dataset contains a total of 100,000 images, evenly divided between real and synthetic (fake) content. These images are organized into ten distinct object classes, such as cats, dogs, horses, and vehicles.

To facilitate the training and evaluation of the classification model, the dataset was split into 80,000 images for training (40,000 real + 40,000 fake) 20,000 images for testing (10,000 real + 10,000 fake).

The real images were sourced from authentic datasets, while the synthetic images were generated using Stable Diffusion v1.4, a powerful Latent Diffusion Model (LDM). All images were resized to 32×32 pixels, maintaining consistent input dimensions for the convolutional neural network.

This balanced dataset supports a binary classification task, where each image is labeled as either "REAL" or "FAKE." Its diversity and scale make it well-suited for training robust models and studying the subtle visual cues that distinguish synthetic images from genuine ones.

5 | Implementation and Results

5.1 | Real Data Collection

The real images for this study were sourced from the Artifact dataset, containing 100,000 32×32-pixel images across ten classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck). 80,000 images were allocated for training, and 20,000 for testing, with the "REAL" label assigned to this dataset.

5.2 | Synthetic Data Generation

Synthetic images were generated using Stable Diffusion v1.4, a publicly available Latent Diffusion Model. Textual prompts, augmented with diverse modifiers, were used to guide the generation process for each class. For example, the airplane class used variations like "fighter jet" or "airplane flying over mountains." 50 denoising steps were applied to generate each synthetic image, initially at 512×512 resolution, later resized to 32×32 pixels to match Artifact dimensions. 50,000 synthetic images were created in total, balanced across the ten classes and split equally between training and testing.

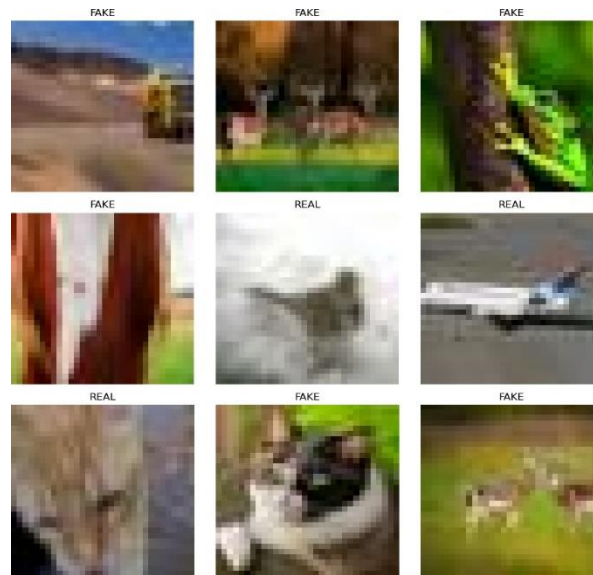


Figure 1. Examples of real and synthetic images from the Artifact dataset.

5.3 | Image Classification

A Convolutional Neural Network (CNN) was designed for binary classification. The architecture included convolutional layers for feature extraction, ReLU activations, pooling operations, and fully connected dense layers leading to a single Sigmoid output neuron.

Extensive hyperparameter tuning was conducted, testing various numbers of convolutional filters and dense neurons. Metrics such as accuracy, precision, recall, and F1-score were used for evaluation, emphasizing the importance of minimizing false positives and false negatives.

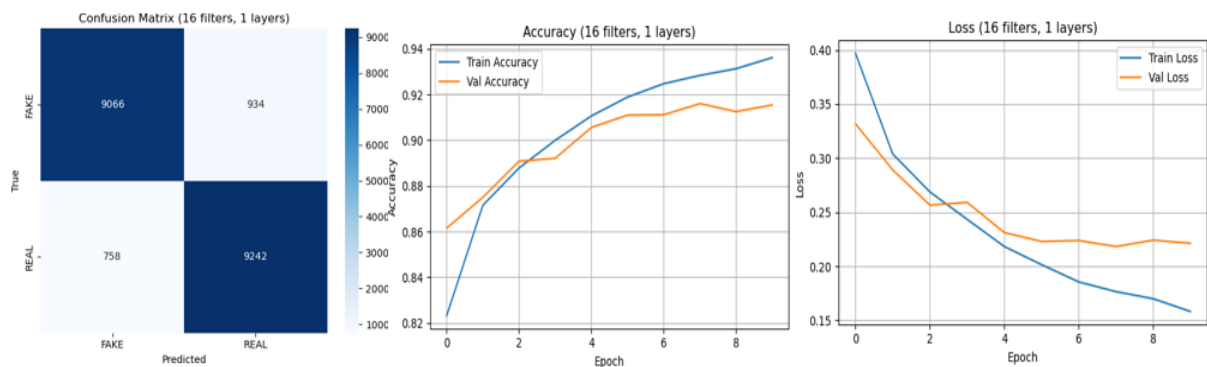


Figure 2. Confusion Matrix and performance metrics (Accuracy, Loss) for CNN with filter 16 and layer 1.

Table 1. Classification report (precision, recall, F1-score, support).

Accuracy (%)	Validation loss	precision	recall	F1-score
91.54	0.2213	0.9155	0.9154	0.9154

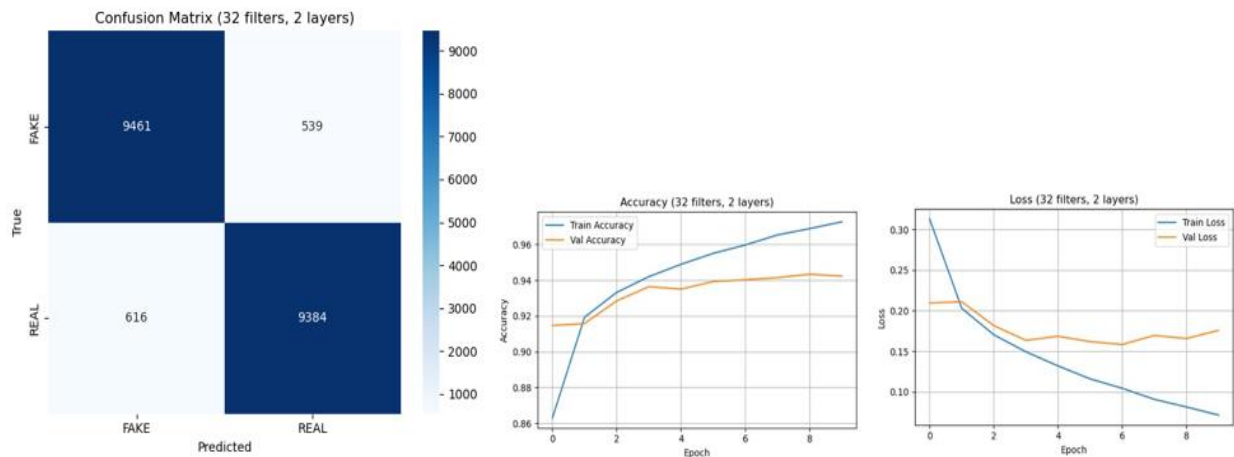


Figure 3. Confusion Matrix and performance metrics (Accuracy, Loss) for CNN with filter 32 and layer 2.

Table 2. Classification report (precision, recall, F1-score, support).

Accuracy (%)	Validation loss	precision	recall	F1-score
94.23	0.1757	0.9423	0.9423	0.9422

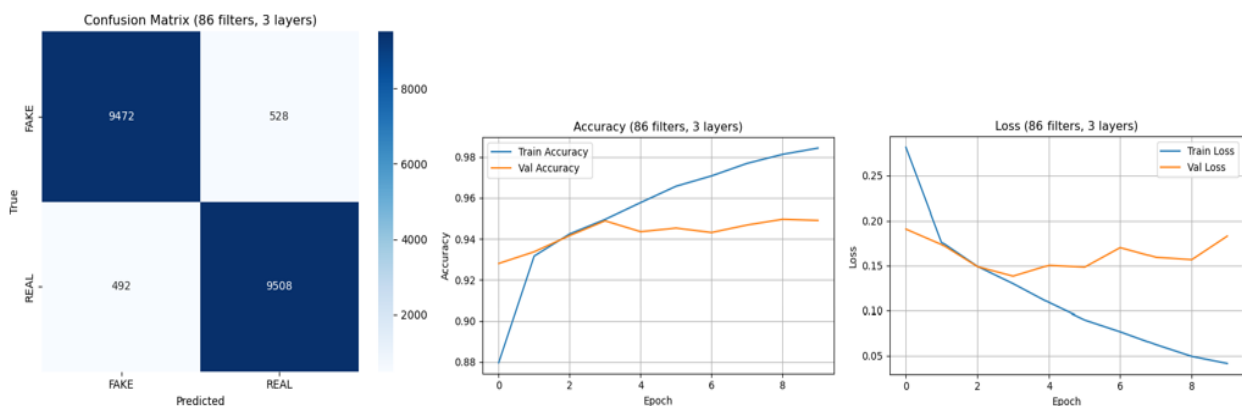


Figure 4. Confusion Matrix and performance metrics (Accuracy, Loss) for CNN with filter 86 and layer 3.

Table 3. Classification report (precision, recall, F1-score, support).

Accuracy (%)	Validation loss	precision	recall	F1-score
94.90	0.1830	0.9490	0.9490	0.9490

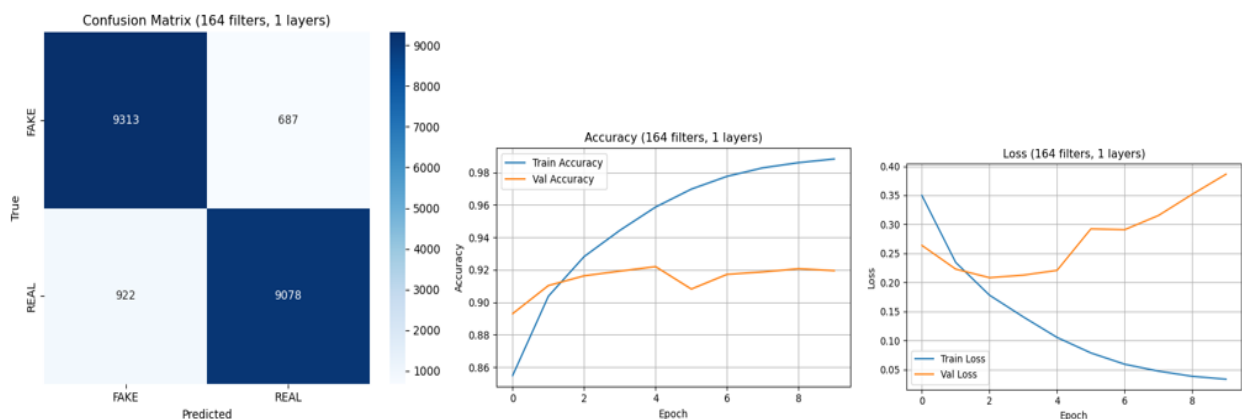


Figure 5. Confusion Matrix and performance metrics (Accuracy, Loss) for CNN with filter 164 and layer 1.

Table 4. Classification report (precision, recall, F1-score, support).

Accuracy (%)	Validation loss	precision	recall	F1-score
91.95	0.3864	0.9198	0.9196	0.9195

Since the "FAKE" class is zero and the "REAL" class is one, the chance of that class is represented by a value that is closer to either of the two values. The values are rounded to the closest value for inference; however, this helps with learning because it is differentiable. While the network's objective is to decrease binary cross-entropy loss through backpropagation, this study also observes an increased number of classification measures. These include the Precision metric, which quantifies the proportion of predictive positive instances that are positive and enables the examination of false positives:

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}}$$

Recall is a measure of how many positive cases are correctly predicted, which enables analysis of false negative predictions:

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives} + \text{False negative}}$$

This measure is particularly important in this case, as it is in fraud detection, since a false negative would falsely accuse the author of generating their image with AI. Finally, the F-1 score is considered:

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

5.4 | Explainable AI

To understand the decision-making process of CNN, Gradient Class Activation Mapping (Grad-CAM) was employed. Grad-CAM visualizes the areas within an input image that the network focuses on when making a classification decision, revealing whether the model bases its decisions on relevant or irrelevant features.

5.5 | Experimental Hardware and Software

All experiments were conducted using TensorFlow, with random seeds fixed for reproducibility. Synthetic images were generated on an Nvidia Apple M1 GPU. The study relied on publicly available models and datasets, ensuring transparency and replicability.

6 | Performance Evaluation

6.1 | Dataset Exploration

The synthetic images generated exhibited high realism, capturing complex visual effects such as depth, reflections, and motion blur. Nonetheless, some imperfections were noted, such as pseudo-textual artifacts and anatomical inaccuracies. Overall, the dataset successfully challenged human and machine capabilities in distinguishing real from synthetic content.

6.2 | Classification Results

Initial CNN configurations achieved strong results, with average accuracies around 92.89%. The best-performing feature extractor used two convolutional layers with 128 filters, reaching 94.30% accuracy.

Table 5. Observed classification accuracy metrics for feature extraction networks.

Lawyer / filter	1	2	3
16	0.9154	0.9412	0.9484
32	0.9163	0.9423	0.9493
64	0.9102	0.9380	0.9490
128	0.9195	0.9430	0.9380

Table 6. Observed validation loss for the filters within the Convolutional Neural Network.

Lawyer / filter	1	2	3
16	0.2213	0.1974	0.1447
32	0.2833	0.1757	0.1459
64	0.3948	0.2765	0.1830
128	0.3864	0.1390	0.1822

Table 7. Observed validation precision for the filters within the Convolutional Neural Network.

Lawyer / filter	1	2	3
16	0.9155	0.9413	0.9484
32	0.9163	0.9423	0.9494
64	0.9131	0.9386	0.9490
128	0.9198	0.9829	0.9388

Table 8. Observed validation recall for the filters within the Convolutional Neural Network.

Lawyer / filter	1	2	3
16	0.9154	0.9412	0.9484
32	0.9163	0.9423	0.9493
64	0.9102	0.9380	0.9490
128	0.9196	0.9505	0.9398

Table 9. Observed validation F1-Score for the filters within the Convolutional Neural Network.

Lawyer / filter	1	2	3
16	0.9154	0.9411	0.9483
32	0.9163	0.9422	0.9493
64	0.9100	0.9380	0.9490
128	0.9195	0.9883	0.9325

7 | Conclusion and Future Work

This study demonstrated a practical approach to enhancing the detection of AI-generated images using computer vision and explainable AI. The creation of the Artifact dataset provides a vital resource for future research efforts, and the CNN-based classification system showed excellent performance.

However, the study also highlighted that synthetic imagery is becoming increasingly difficult to detect, especially as generative models continue to evolve.

Integrating attention-based models like Vision Transformers for even more precise classification. Extending detection efforts to domains such as human faces, medical imaging, and satellite photos. Combining visual detection with metadata analysis for more robust authentication systems.

Exploring ethical frameworks for deploying detection systems in real-world applications.

In conclusion, this study contributes to the vital task of preserving trust in visual media in the era of advanced AI synthesis, offering both practical tools and theoretical insights for future interdisciplinary research.

Funding

This research has no funding source.

Data Availability

The datasets generated during and/or analyzed during the current study are not publicly available due to the privacy-preserving nature of the data, but are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that there is no conflict of interest in the research.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

- [1] Gawali, V., Chaubey, C., Gaikwad, M., Gidde, A., & Bhelkar, N. (2024). Study on AI-generated fake-media detection. *International Research Journal on Advanced Engineering and Management*, 2(10), 3181–3185.
- [2] Park, D., Na, H.-S., & Choi, D. (2024). Performance comparison and visualization of AI-generated-image detection methods. *IEEE Access*. <https://doi.org/10.1109/ACCESS.2024.3394250>
- [3] Roose, K. (2022, September 2). An AI-generated picture won an art prize. Artists aren't happy. *The New York Times*. <https://www.nytimes.com/2022/09/02/technology/ai-artificial-intelligence-artists.html>
- [4] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10684–10695). <https://doi.org/10.1109/CVPR52688.2022.01042>
- [5] Deb, D., Zhang, J., & Jain, A. K. (2020). AdvFaces: Adversarial face synthesis. In *2020 IEEE International Joint Conference on Biometrics (IJCB)* (pp. 1–10). IEEE. <https://doi.org/10.1109/IJCB48548.2020.9304898>
- [6] Ara, A., Alam, M. S., & Mifa, K. A. F. (2024). A comparative review of AI-generated image detection across social media platforms. *Global Mainstream Journal of Innovation, Engineering & Emerging Technology*, 3(1), 2834–2739. *Global Mainstream Journal prefix DOI: 10.62304*
- [7] Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, 25(5), 388–402.
- [8] Bonettini, N., Bestagini, P., Milani, S., & Tubaro, S. (2021). On the use of Benford's law to detect GAN-generated images. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 5495–5502). IEEE. <https://doi.org/10.1109/ICPR48806.2021.9412944>

Disclaimer/Publisher's Note: The perspectives, opinions, and data shared in all publications are the sole responsibility of the individual authors and contributors, and do not necessarily reflect the views of Sciences Force or the editorial team. Sciences Force and the editorial team disclaim any liability for potential harm to individuals or property resulting from the ideas, methods, instructions, or products referenced in the content.