International Journal of Science, Strategic Management and Technology
Volume **02** Issue **05** May-**2026** | ISSN: **3108-1762 (Online)** | Impact Factor: 3.8
An International, Peer-Reviewed, Open Access Scholarly Journal Indexed in recognized academic databases

Synthetic Image Detection using Deep Learning

[1] P. Marimuthu
[2] B. Parameshwaran
[3] R. A. Anusurya
UG Student,
Vels Institute of Science,
Technology and Advanced Studies (VISTAS),
Pallavaram, Chennai - 600117,
Tamil Nadu, India.

[4] Dr. A. Akila
Associate Professor,
Vels Institute of Science,
Technology and Advanced Studies (VISTAS),
Pallavaram, Chennai - 600117,
Tamil Nadu, India.

https://doi.org/10.55041/ijsmt.v2i5.015

Cite this Article: Marimuthu, P., Parameshwaran, B. & Anusurya, R. A. (2026). Synthetic Image Detection using Deep Learning. International Journal of Science, Strategic Management and Technology, *02*(05). https://doi.org/10.55041/ijsmt.v2i5.015

ABSTRACT—The proliferation of AI-generated synthetic images produced by Generative Adversarial Networks (GANs) and diffusion-based models has created critical challenges for digital forensics, media authentication, and public trust. This paper presents a complete Synthetic Image Detection (SID) system capable of classifying any digital image as REAL or AI-generated without requiring access to any external dataset at inference time. The system combines two complementary deep learning architectures — a fine-tuned ResNet50 Convolutional Neural Network and a Vision Transformer (ViT-B/16) — both trained on the CIFAKE benchmark comprising 120,000 images. ResNet50 achieved 98.08% test accuracy with AUC-ROC of 0.9971, and ViT-B/16 achieved 98.05% accuracy with AUC-ROC of 0.9984, both substantially exceeding the published CIFAKE baseline of 92.98% (Bird and Lotfi, 2024). GradCAM++ (Gradient-weighted Class Activation Mapping++) was implemented from scratch to generate pixel-level visual explanations identifying the specific image regions influencing each classification decision. Beyond deep learning, a novel multi-signal forensic analysis pipeline was developed incorporating six independent pixel-level detectors: Discrete Fourier Transform checkerboard analysis, Error Level Analysis, SRM noise residual inspection, colour channel statistical analysis, Local Binary Pattern texture entropy, and Haar wavelet energy ratio computation. All six signals are combined with the two deep learning models in a weighted mega-ensemble producing a final verdict purely from the uploaded image pixels. The complete system is deployed as a RESTful API using FastAPI with a five-page Streamlit web interface providing real-time detection, GradCAM++ visualisation, and forensic signal breakdown.

KEYWORDS—*Synthetic image detection, deep learning, ResNet50, Vision Transformer, ViT-B/16, GradCAM++, CIFAKE, forensic signal analysis, Error Level Analysis, DFT, image forensics, binary classification, FastAPI, Streamlit.*

I. INTRODUCTION

The rapid advancement of generative artificial intelligence has fundamentally altered the landscape of digital media. Generative Adversarial Networks (GANs), introduced by Goodfellow et al. in 2014, established a framework for synthesising photorealistic images by training a generator network to deceive a discriminator, producing outputs increasingly indistinguishable from genuine photographs. Subsequent architectures including ProGAN, StyleGAN2, and BigGAN achieved commercial-quality image generation, while the emergence of diffusion-

based models — most notably Stable Diffusion, DALL·E 3, and Midjourney — further democratised AI image synthesis by enabling text-to-image generation accessible to the general public.

The societal consequences of this technological progress are significant. AI-generated images have been weaponised for misinformation campaigns, identity fraud, non-consensual imagery, and interference in democratic processes. Human observers demonstrate an inability to reliably distinguish AI-generated faces and scenes from real photographs, with reported accuracy rates only marginally above random chance (~53%). This fundamental limitation of human perception motivates the development of automated computational systems capable of accurately and efficiently detecting synthetic images at scale.

This project addresses these gaps by building a complete end-to-end Synthetic Image Detection system. Two state-of-the-art deep learning models are trained and evaluated on the CIFAKE benchmark, GradCAM++ explainability is implemented from scratch, six independent forensic signal detectors are developed, and the complete pipeline is deployed as a web application. The system classifies any uploaded image as REAL or FAKE without requiring access to any external dataset or GPU at inference time for the forensic signal analysis pathway.

II. LITERATURE REVIEW

Bird and Lotfi (2024) introduced the CIFAKE dataset — a benchmark comprising 120,000 images — and demonstrated that a CNN trained on this data achieved 92.98% classification accuracy using Gradient Class Activation Mapping (Grad-CAM) for explainability. Their analysis revealed that the model focused on background artefacts rather than foreground objects, establishing a critical interpretability baseline for the field [1].

Corvi et al. (2023) examined forensic traces left by diffusion models and demonstrated that detectors trained on GAN images generalise poorly to diffusion-generated content, as the two generation paradigms produce distinct frequency-domain artefacts. This cross-architecture generalisation problem motivates the use of architecturally diverse datasets and pixel-level signal analysis [2].

Selvaraju et al. (2017) introduced Grad-CAM, which generates class-discriminative visualisations through gradient-weighted feature map averaging. Chattopadhay et al. (2018) subsequently proposed GradCAM++ with second-order gradient alpha weights, providing improved localisation for distributed artefacts — a critical enhancement for forensic applications where multiple small background regions jointly influence classification [3].

Dosovitskiy et al. (2021) introduced the Vision Transformer (ViT), demonstrating that patch-based self-attention mechanisms achieve competitive image classification performance without convolution. Subsequent work by Das et al. (2025) applied a ViT with edge-based variance modules to CIFAKE, achieving 97.75% accuracy and establishing the Transformer as a viable architecture for synthetic image detection [4].

Zhang et al. (2019) demonstrated that CNN-generated images exhibit characteristic spectral discrepancies in the frequency domain exploitable by ResNet-based binary classifiers. This foundational frequency-domain insight directly informed the design of the DFT checkerboard signal detector in the present work [5].

III. PROBLEM DEFINITION

Synthetic image detection presents several interrelated technical challenges. First, achieving classification accuracy exceeding published state-of-the-art results on the CIFAKE benchmark requires careful hyperparameter tuning and training strategy design. Second, detection decisions must be explainable and interpretable — the 'black box' nature of deep learning classifiers is unacceptable in forensic and legal contexts. Third, the system must be capable of classifying arbitrary uploaded images without access to training data at inference time. Fourth, the system must provide a user-accessible interface presenting detection results, confidence scores, visual explanations, and forensic signal breakdowns in a format interpretable by non-expert users.

A further challenge is the domain gap between training data and real-world deployment. The CIFAKE dataset comprises 32×32 pixel CIFAR-10 style images upscaled to 224×224, while real-world photographs are typically high-resolution with complex photographic characteristics. This creates a distribution shift that reduces single-model reliability and motivates the addition of model-independent forensic signals capable of detecting artefacts based on mathematical properties rather than learned feature representations.

IV. SYNTHETIC IMAGE DETECTION SYSTEM

The Synthetic Image Detection system comprises seven integrated components forming a complete detection pipeline from image upload to final verdict:

A. Dataset and Preprocessing

The CIFAKE dataset comprises 120,000 balanced images: 60,000 authentic CIFAR-10 photographs (REAL) and 60,000 AI-generated counterparts produced using Stable Diffusion version 1.4 (FAKE). The dataset is split into 85,000 training images (50%), 15,000 validation images (12.5%), and 20,000 test images (16.7%). All images are resized to 224×224 pixels and normalised using ImageNet mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225]. Training images undergo augmentation including RandomHorizontalFlip, RandomRotation (±15°), ColorJitter, and GaussianBlur (p=0.2) to improve generalisation.

B. ResNet50 Deep Learning Model

A ResNet50 Convolutional Neural Network pretrained on ImageNet-1K (V2 weights) is fine-tuned on CIFAKE. The pretrained classification head is replaced with a custom binary detection head comprising Dropout(0.4), Linear(2048, 512), ReLU, Dropout(0.3), and Linear(512, 2). Training employs the AdamW optimiser with learning rate 5e-5 and weight decay 1e-4, cosine annealing learning rate scheduling, label smoothing of 0.05, and early stopping with patience 5. The first four layer groups are frozen (freeze_until=4), enabling 94.1% of parameters (23.1M of 24.5M) to train on CIFAKE. Mixed-precision training via PyTorch's autocast reduces VRAM requirements on the GTX 1050 Ti GPU.

C. Vision Transformer (ViT-B/16)

The ViT-B/16 model pretrained on ImageNet-21k is fine-tuned using the google/vit-base-patch16-224 checkpoint. Each 224×224 image is divided into 196 non-overlapping 16×16 pixel patches, linearly projected to 768-dimensional embeddings, and processed by 12 Transformer encoder blocks with 12-head self-attention. A custom classification head (LayerNorm, Dropout(0.3), Linear(768, 256), GELU, Dropout(0.2), Linear(256, 2)) replaces the original 1000-class head. Training uses AdamW with learning rate 2e-5, weight decay 1e-2, cosine scheduling with 3,985 linear warm-up steps, and gradient clipping at max_norm=1.0.

D. GradCAM++ Explainability

GradCAM++ is implemented from scratch using PyTorch forward and backward hooks registered on ResNet50's layer4. For each prediction, the system computes the alpha-weighted linear combination of feature maps, applies ReLU, upsamples to 224×224, and normalises to [0,1]. The resulting heatmap is overlaid on the input image using a jet colourmap (alpha=0.45), producing four visual outputs: the original image, the raw heatmap, the blended overlay, and a key-region mask (threshold 0.55). The second-order gradient alpha formula is:

$$\alpha_k = (\partial^2 Yc/\partial A^2_k) / [2(\partial^2 Yc/\partial A^2_k) + \Sigma_{ij} A_{ij\_k}(\partial^3 Yc/\partial A^3_k)].$$

E. Six Forensic Signal Detectors

Six independent pixel-level forensic signals are computed from the raw image bytes without any trained model, external dataset, or internet connection. Each signal returns a fake_score in [0,1], a label (REAL/FAKE/UNCERTAIN), and detailed statistics:

- DFT Checkerboard (weight 0.22): Computes 2D FFT magnitude spectrum and measures power in the GAN upsampling frequency band (30–45% of maximum radius) relative to total spectral power.
- ELA Analysis (weight 0.22): Re-saves the image at JPEG quality 92 and measures pixel-wise differences. AI-generated images show abnormally high error levels (ela_mean 15–40) vs real photos (2–8).
- SRM Noise Residual (weight 0.16): Applies the Laplacian filter [[0,-1,0],[-1,4,-1],[0,-1,0]]/4 to isolate noise residuals and measures smoothness, standard deviation, and periodic frequency content.
- Colour Channel Statistics (weight 0.16): Computes inter-channel Pearson correlation coefficients and saturation statistics. AI generators exhibit avg_corr > 0.90 due to shared feature map representations.
- LBP Texture Entropy (weight 0.12): Computes 8-neighbour Local Binary Pattern codes and measures Shannon entropy of the LBP histogram. Low entropy (< 6.0) indicates unnaturally repetitive texture.
- Wavelet Energy Ratio (weight 0.12): Applies 3-level Haar decomposition and measures energy ratios across approximation and detail sub-bands. Non-1/f energy ordering indicates generative artefacts.

The six signals are combined into a weighted ensemble score, producing an overall verdict with confidence level.

F. Mega-Ensemble Combination

The final verdict combines all detection sources using a weighted scheme: ResNet50 fake probability (35%), ViT-B/16 fake probability (35%), and the six-signal ensemble score (30%). Dynamic weight rebalancing is applied when the two deep learning models disagree — ViT weight is increased to 45% and ResNet50 reduced to 25%, reflecting ViT's superior generalisation on novel generator architectures. A decision threshold of 0.50 is used for FAKE verdicts, with UNCERTAIN assigned for scores between 0.45 and 0.50.

G. Deployment

The complete system is deployed as a FastAPI REST API (version 0.110.0) served by uvicorn, exposing eight endpoints: GET /health, GET /metrics, POST /predict/resnet, POST /predict/vit, POST /predict/compare, POST /predict, POST /predict/signals, and POST /predict/ensemble. A five-page Streamlit web frontend (version 1.32.0) at localhost:8501 provides: an image detection page with GradCAM++ heatmap, a forensic analysis page showing all six signal gauge charts and ensemble breakdown, a model metrics comparison page, an API status page, and an about page.

V. RESULTS AND DISCUSSION

Both models were evaluated on the held-out CIFAKE test set of 20,000 images not seen during training. Table I presents the key performance metrics.

TABLE I. PERFORMANCE COMPARISON ON CIFAKE TEST SET

Model	Accuracy	AUC-ROC	F1 Score	Train Time
Bird & Lotfi (2024)	92.98%	—	—	—
Our ResNet50	98.08%	0.9971	0.9808	777 min
Our ViT-B/16	98.05%	0.9984	0.9805	2,147 min

ResNet50 achieved the highest accuracy (98.08%) with early stopping at epoch 12 of 20. The confusion matrix reveals 9,882 correct FAKE classifications (98.82% recall) and 9,735 correct REAL classifications (97.35% recall). The asymmetry reflects the forensically desirable property that FAKE images are more reliably detected — false negatives (missed FAKE images) represent the more serious error type in a detection context.

ViT-B/16 achieved near-identical accuracy (98.05%) with a marginally higher AUC-ROC (0.9984 vs 0.9971), indicating better discrimination at all classification thresholds. The ViT's global self-attention mechanism captures image-wide consistency artefacts distributed across multiple patches simultaneously, while ResNet50's convolutional layers detect local texture anomalies. Both architectures exceed all published CIFAKE benchmarks, including the CIFAKE baseline (92.98%), SE-ResNet50 (96.12%), and PV-ISM ViT (96.60%).

GradCAM++ analysis confirmed that the model focuses on background texture regions in FAKE images rather than foreground objects — consistent with the findings of Bird and Lotfi (2024). Layer4 activations showed the clearest forensic patterns, with FAKE images exhibiting concentrated heatmap regions over background artefacts and REAL images showing distributed activations over natural scene structures.

Table II presents the six forensic signal characteristics.

TABLE II. FORENSIC SIGNAL CHARACTERISTICS AND WEIGHTS

Signal	Detects	Weight	Shade
DFT Checkerboard	GAN upsampling periodic frequency artefacts	0.22	Light
ELA Analysis	JPEG compression history inconsistency	0.22	White
SRM Noise	Unnatural noise residual patterns	0.16	Light
Colour Stats	RGB inter-channel correlation anomalies	0.16	White
LBP Texture	Micro-texture Shannon entropy deficit	0.12	Light

Wavelet Energy	Multi-scale Haar energy ratio anomaly	0.12	White

The DFT and ELA signals demonstrated the highest individual reliability. No single signal achieved deep learning model accuracy independently, confirming the primary design insight: deep learning feature extraction is essential, and the six signals provide supplementary transparent evidence rather than a standalone classification pathway.

VI. CONCLUSION

This paper presented a complete Synthetic Image Detection system combining fine-tuned ResNet50 and ViT-B/16 deep learning models with six independent pixel-level forensic signal detectors and GradCAM++ explainability, deployed as a full-stack web application. Both deep learning models achieved approximately 98% test accuracy on the CIFAKE benchmark — a 5.10 percentage point improvement over the published baseline — through careful training strategy design including layer-selective unfreezing, cosine annealing scheduling, and label smoothing. The near-identical performance of CNN and Transformer architectures is a significant empirical contribution, challenging the assumption of Transformer superiority at standard benchmark scale.

The six forensic signals provide model-independent, mathematically transparent supplementary detection evidence that remains valid for generator architectures outside the training distribution. The complete system is practically deployable on commodity GPU hardware (GTX 1050 Ti, 4.3 GB VRAM) and accessible via a browser interface. Future work includes training on higher-resolution diverse datasets (e.g. GenImage), multi-class source attribution, adversarial robustness training, and browser extension deployment for real-time web image analysis.

REFERENCES

[1] J. J. Bird and A. Lotfi, "CIFAKE: Image classification and explainable identification of AI-generated synthetic images," IEEE Access, vol. 12, pp. 15642–15650, 2024.

[2] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," in Proc. IEEE ICASSP, 2023.

[3] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalised gradient-based visual explanations for deep convolutional networks," in Proc. IEEE WACV, 2018.

[4] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.

[5] X. Zhang, S. Karaman, and S. F. Chang, "Detecting and simulating artifacts in GAN fake images," in Proc. IEEE WIFS, 2019.

[6] 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 ICCV, 2017.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE CVPR, pp. 770–778, 2016.

[8] I. Goodfellow et al., "Generative adversarial nets," in Advances in Neural Information Processing Systems (NeurIPS), vol. 27, 2014.

[9] R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, "Intriguing properties of synthetic images: From GANs to diffusion models," in Proc. IEEE/CVF CVPRW, 2023.

[10] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, "Raising the bar of AI-generated image detection with CLIP," in Proc. IEEE/CVF CVPR, pp. 4356–4366, 2024.

[11] H. Liu, Z. Tan, C. Tan, Y. Wei, J. Wang, and Y. Zhao, "Forgery-aware adaptive transformer for generalizable synthetic image detection," in Proc. IEEE/CVF CVPR, 2024.