iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

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

DOI: https://doi.org/10.22214/ijraset.2026.85329

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue X Oct 2026- Available at www.ijraset.com

Uthentica AI: An Explainable Multimodal System for Synthetic Media Detection Using Deep Features, Reconstruction Error and Frequency-Domain Forensics

Amaresh Kannan K[1], Aravindh NK[2], Mrs. Sindhiya R[3]
Department of Computer Science and Engineering, K.L.N. College of Engineering (An Autonomous Institution, Affiliated to Anna University, Chennai), Tamil Nadu, India

Abstract: ***The rapid improvement of generative models has made AI-generated images, cloned voices and manipulated video increasingly difficult to distinguish from authentic media, while most deployed detectors examine a single modality and return an opaque score. This paper presents Authentica AI, a web-based multimodal synthetic-media detection system that automatically routes each upload to a modality-specific detector. Images are analysed by a pretrained convolutional backbone (EfficientNet-B0, with a ResNet-50 variant), a diffusion-reconstruction error map, and five handcrafted forensic signals — spectral energy loss, spatial variation, edge-sharpness consistency, colour-channel correlation and local-variance entropy — which are fused by a weighted ensemble and a calibrated sigmoid into a single synthetic-probability score. Audio is converted to log-Mel spectrograms and classified by a ResNet-18 model. Every image verdict is accompanied by a forensic heatmap that localises the regions that drove the decision, and results are stored in a PostgreSQL database and presented through an interactive dashboard with batch processing. In a prototype evaluation on a five-image batch the system returned verdicts with a mean latency of 263.3 ms per image on a CPU, and per-stage profiling of a single run (288.1 ms) shows that the reconstruction stages account for about two thirds of the processing time. The paper also states the limits of this evaluation: no labelled benchmark accuracy is reported, and large-scale benchmarking and video support are identified as future work.***

Keywords: deepfake detection; synthetic media; multimodal forensics; explainable AI; diffusion reconstruction error; frequency-domain analysis; audio spectrogram classification.

I. INTRODUCTION

Generative Adversarial Networks (GANs) [9] and diffusion models [8] can now produce photorealistic images, cloned human voices and manipulated video with little technical effort. These techniques have legitimate creative and commercial uses, but they also enable synthetic media, commonly called deepfakes, that can spread misinformation, impersonate individuals and erode public trust in digital content. As generators improve, the artefacts that once made fake content easy to spot, such as unnatural blinking, inconsistent lighting or robotic speech, are steadily disappearing, which creates a need for automated, reliable and interpretable detection that ordinary users can operate.

Existing detectors have three recurring weaknesses. First, many are trained and tested on one modality and so miss content in which, for example, a synthetic voice accompanies an unaltered video. Second, detectors that perform well on the generators seen during training often degrade on newly released generators and on media that has been compressed or re-encoded by social platforms [2, 13]. Third, most return a bare label or probability, giving the user no way to see *why* a file was flagged.

Authentica AI addresses these weaknesses in a single platform. The main contributions of this work are:

1) A media-type router that sends each upload to a detector built for that modality (image or audio in the current implementation; video is planned), instead of using one generic classifier.

2) A hybrid image pipeline that combines deep convolutional features, a diffusion-reconstruction error map and five handcrafted frequency-domain and statistical forensic signals, fused by a weighted ensemble and calibrated sigmoid into a single synthetic-probability score.

3) Region-level explanations: a forensic heatmap overlay for every analysed image, so that each verdict can be inspected rather than taken on trust.

©IJRASET: All Rights are Reserved

4) A working web platform with a confidence gauge, metrics panel, batch analysis, per-stage execution trace and a persistent result store, together with a transparent report of its measured latency and of the limits of the present evaluation.

The remainder of the paper is organised as follows. Section 2 reviews related work, Section 3 describes the proposed system, Section 4 gives implementation details, Section 5 reports prototype results, Section 6 discusses limitations, and Section 7 concludes with future work.

II. RELATED WORK

A. *Multimodal and survey literature*

Raza and Malik [1] propose a unified audiovisual deepfake detector that uses a ResNet-1D for audio and a 3D-ResNet for video, with independent and joint modality-mixing layers feeding a multilabel classification head. Trained on FakeAVCeleb and evaluated on further datasets, it reaches 92.9% accuracy on FakeAVCeleb and outperforms unimodal, multimodal and ensemble baselines, and its multilabel output gives separate audio and video predictions. The authors note that robustness to noise and to laundering scenarios needs further study. Lee et al. [3] combine audio, visual and audio-visual streams with one-class learning (OC-Softmax) to improve robustness to unseen attacks; their approach still degrades on subtle manipulations and struggles with audio-visual desynchronisation. Lin et al. [2] survey detection of content produced by large AI models across text, image, video, audio and multimodal data, propose a taxonomy that separates “pure detection” from “beyond detection” and name robustness to noisy social-media content as an open challenge. Together these works motivate combining modalities and designing for robustness, but none of them provides a region-level explanation for a lay user.

B. *Image forensics and generator artefacts*

Wang et al. [13] showed that a CNN classifier trained on images from a single GAN can generalise surprisingly well to other generators when suitable augmentation is used, although generalisation is not guaranteed. Frequency-domain analysis has been a productive direction: Durall et al. [10] and Frank et al. [11] showed that up-sampling in generative networks leaves characteristic spectral artefacts, which motivates the spectral signals used here. For diffusion-generated images, Wang et al. [7] introduced DIRE, which measures the error between an image and its reconstruction by a pretrained diffusion model; real images tend to be reconstructed less faithfully than diffusion-generated ones, so the error acts as a detection cue. Authentica AI adopts this reconstruction-error idea as one input to its fused score. For explanation, gradient-based visualisation such as Grad-CAM [12] is widely used; Authentica AI currently uses an edge-filtered forensic heatmap and lists Grad-CAM as planned work.

C. *Positioning*

Table 1 summarises how the proposed system relates to the multimodal works reviewed above.

Table 1. Comparison with related multimodal detection work.

Work	Modalities	Approach	Limitation noted
Raza & Malik [1]	Audio + video	ResNet-1D / 3D-ResNet, modality-mixing layers, multilabel head	Noise and laundering robustness unverified
Lee et al. [3]	Audio + visual	Multi-stream fusion with one-class learning (OC-Softmax)	Subtle manipulations; desynchronisation
Lin et al. [2]	Text, image, video, audio	Survey and taxonomy (no new model)	Robustness to noisy social-media content is open
Authentica AI (this work)	Image + audio (video planned)	Router; CNN features + reconstruction error + 5 forensic signals; spectrogram ResNet-18; heatmap	No large-scale labelled benchmark yet (Section 6)

III. PROPOSED SYSTEM

A. *Overview and media-type routing*

A user uploads a file through the web interface. A preprocessing and media-type router identifies whether the file is an image (JPG, PNG, WebP) or audio (WAV, MP3, M4A) and sends it to the corresponding detector. The image detector and the audio detector each produce a prediction; a result aggregator and confidence-scoring stage then builds the final authenticity report, containing the real/fake verdict, the confidence score and supporting visual evidence.

Figure 1 shows the processing architecture and Figure 2 the data flow.

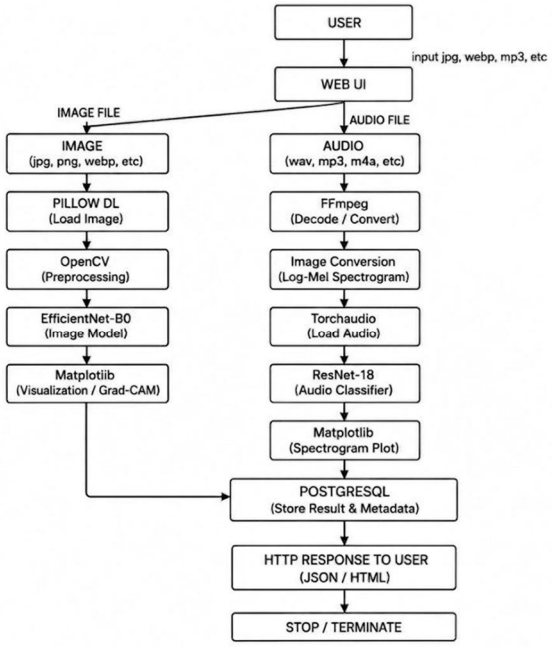


Fig. 1. System architecture. Images pass through Pillow, OpenCV and an image model; audio is decoded with FFmpeg, converted to a log-Mel spectrogram and classified by ResNet-18. Both branches write results and metadata to PostgreSQL before an HTTP response is returned.

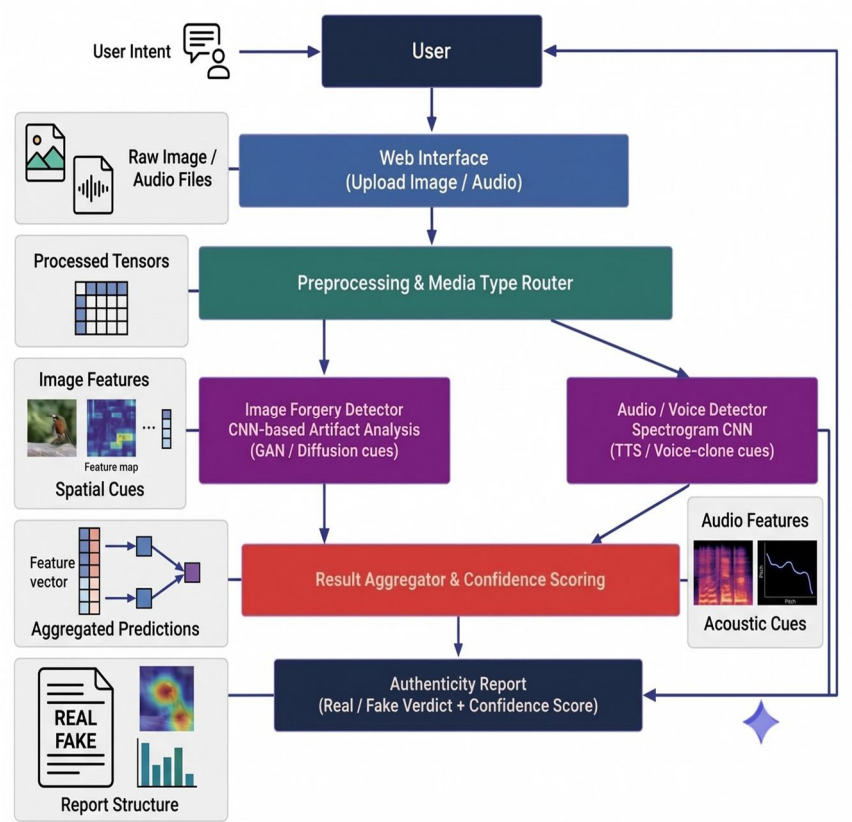


Fig. 2. Data flow from user intent to the final authenticity report, through the media-type router, the image and audio detectors, and the result aggregator.

B. Image detection pipeline

Preprocessing. Each image is decoded and validated, resized to 256×256, centre-cropped to 224×224 and normalised, matching the input distribution expected by the pretrained backbone. A one-time warm-up pass removes cold-start latency before real inference.

Deep feature classification. A pretrained convolutional backbone, EfficientNet-B0 [4] with its final layer replaced by a two-class head (a ResNet-50 variant [5] is also supported), outputs a synthetic-class probability s_0. Custom-trained weights are used when available; otherwise the system falls back to the pretrained model in a demonstration mode.

Diffusion-reconstruction error. Following the idea of DIRE [7], the image is perturbed and reconstructed by a diffusion model (a DDPM forward pass at noise level t = 200 with U-Net denoising in the implementation described here), and a pixel-wise squared-error map is computed between the input x and its reconstruction x_{rec}:

$$E(i,j) = \| x(i,j) - x_{\text{rec}}(i,j) \|_2^2 \tag{1}$$

Because generated images tend to lie closer to the generator's manifold, they are expected to be reconstructed with lower error than real photographs; the error map is both a scoring input and the basis for the explanation overlay.

Handcrafted forensic signals. The image is decomposed into five independent signals computed on the grayscale or colour channels: (i) *spectral energy loss*, derived from the log-magnitude of the 2-D Fourier spectrum

$$M(u,v) = 20 \log(|\mathcal{F}\{I_g\}(u,v)| + \varepsilon) \tag{2}$$

where I_g is the grayscale image and □ a small constant, which targets the abnormal high-frequency energy distribution that diffusion models and GAN up-sampling leave behind [10, 11]; (ii) *spatial variation* and (iii) *edge-sharpness consistency*, which capture the unnaturally uniform texture of many generated images (edge sharpness is measured with the variance of the Laplacian); (iv) *colour-channel correlation*; and (v) *local-variance entropy*, which reveal over-smooth statistical patterns that are largely absent in real photographs.

Fusion and calibration. The deep-model probability s_0 and the normalised forensic signals s_1, …, s_K are combined with non-negative weights w_k and passed through a calibrated sigmoid:

$$S = \sigma\left(\alpha \sum_{k=0}^{K} w_k s_k + \beta \right), \qquad \sum_{k=0}^{K} w_k = 1 \tag{3}$$

where □ and □ are calibration parameters and S in [0, 1] is the synthetic probability. In the reference implementation excerpt the deep-model term carries weight 0.6 and the spectral and edge terms 0.25 and 0.15 respectively. An image is labelled *AI-generated* when S exceeds 0.5 and *authentic* otherwise, and a risk level (for example “moderate synthetic risk”) is reported alongside the percentage. Returning a calibrated score rather than a hard decision lets the user judge how confident the system is.

Forensic heatmap. Edge-detection filtering isolates structural inconsistencies introduced during image generation, and the resulting map is rendered with a jet colour map and blended with the original image. This gives a region-level view of the evidence behind the verdict and is produced per image in real time as part of the prediction pipeline.

C. Audio detection pipeline

Uploaded audio is decoded and converted with FFmpeg, transformed into a log-Mel spectrogram and loaded with Torchaudio. A ResNet-18 classifier [5] processes the spectrogram image to identify the prosodic irregularities and spectral discontinuities that are characteristic of AI text-to-speech and voice-cloning systems. Such artefacts are often difficult to hear but visible in the time-frequency domain. The spectrogram plot is rendered alongside the verdict.

D. Web platform and persistence

All detectors are exposed through one browser-based dashboard. A circular gauge shows the synthetic-probability percentage (green for authentic, red for AI-generated); the heatmap is shown beside the original image; and a metrics panel reports the verdict, confidence, risk level, inference latency and media dimensions. A batch mode lists filename, verdict, synthetic percentage and latency for each file, a history view keeps past analyses, and a backend-process view exposes the execution steps of a run. Results and metadata are stored in a PostgreSQL database.

IV. IMPLEMENTATION

The system is implemented in Python 3.11+ with PyTorch and Torchvision for the image models, Torchaudio for audio loading, OpenCV and Pillow for image processing, NumPy, Pandas, scikit-learn, XGBoost and Matplotlib for numerical work and plotting, FFmpeg for audio decoding, and PostgreSQL for storage. Development used Visual Studio Code. Table 2 lists the platform requirements.

Table 2. Hardware and software requirements.

Component	Specification
CPU	Multi-core processor (AMD Ryzen or Intel i5/i7)
Memory / storage	16 GB RAM (minimum); SSD, 256 GB or larger
Operating system	Windows 11 / Windows Server
Language and libraries	Python 3.11+; PyTorch, Torchvision, Torchaudio, OpenCV, Pillow, NumPy, Pandas, scikit-learn, XGBoost, Matplotlib
Media and data	FFmpeg; PostgreSQL

V. PROTOTYPE RESULTS

A. Evaluation Setting

The prototype was exercised through its web dashboard on a CPU-only configuration, using a small set of authentic and AI-generated images and audio clips. The quantitative observations reported below come from the dashboard output for images; ground-truth labels were not formally recorded for this demonstration, so **no accuracy, precision, recall or AUC figure is claimed** in this paper.

B. Batch Analysis

Table 3 and Figure 3 show a five-file batch run. The first file's name indicates that it was exported from a generative-AI tool; the last three are JPEG files received through a messaging application, which also exposes the system to platform recompression. The two files flagged as AI-generated scored 61.9% and 62.0% synthetic probability, and the three files labelled authentic scored between 28.1% and 35.5%, a separation of about 26 percentage points around the 50% decision threshold. Mean latency over the batch was 263.3 ms per image (median 236.6 ms, range 200.7–400.2 ms).

Table 3. Batch results reported by the dashboard (five images, CPU).

#	File	Verdict	Synthetic %	Latency (ms)
1	Gemini_Generated_Image_….png	AI-generated	61.88	400.15
2	Screenshot 2026-09-18 073749.png	AI-generated	62.01	200.70
3	WhatsApp Image 2026-09-01 … PM(1).jpeg	Authentic	35.48	257.09
4	WhatsApp Image 2026-09-01 … PM(2).jpeg	Authentic	28.14	221.80
5	WhatsApp Image 2026-09-01 … PM.jpeg	Authentic	34.21	236.59

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

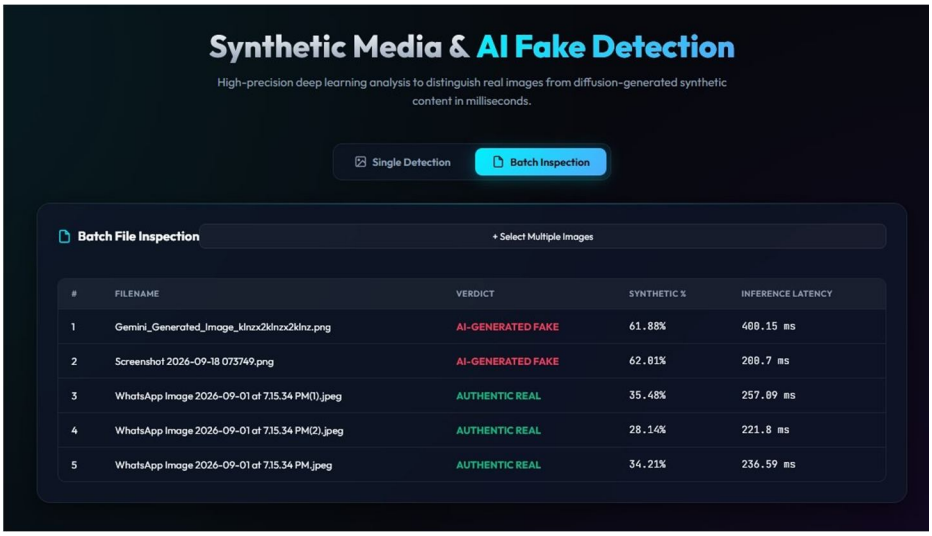


Fig. 3. Batch inspection view showing filename, verdict, synthetic percentage and inference latency per file.

C. Per-stage latency

The backend-process view records the time taken by each stage of a single run. Table 4 and Figure 4 show a total of 288.1 ms. The diffusion-reconstruction pass is the most expensive stage at 135.7 ms (47.1% of the run), and together with the error-map computation (56.6 ms) the reconstruction stages account for 192.3 ms, or 66.7%. Deep feature classification takes 45.2 ms (15.7%) and heatmap rendering 25.5 ms (8.9%); decoding, preprocessing and result compilation together take 25.1 ms. Interactive-speed analysis on a CPU is therefore feasible, and the reconstruction pass is the main target for optimisation.

Table 4. Execution time per pipeline stage for one image (CPU).

Stage	Time (ms)	Share (%)
Image decode and validation	0.8	0.3
Image preprocessing	19.8	6.9
Diffusion reconstruction pass	135.7	47.1
Reconstruction error map	56.6	19.6
Deep feature classification	45.2	15.7
Forensic heatmap render	25.5	8.9
Result compilation	4.5	1.6
Total	288.1	100.0

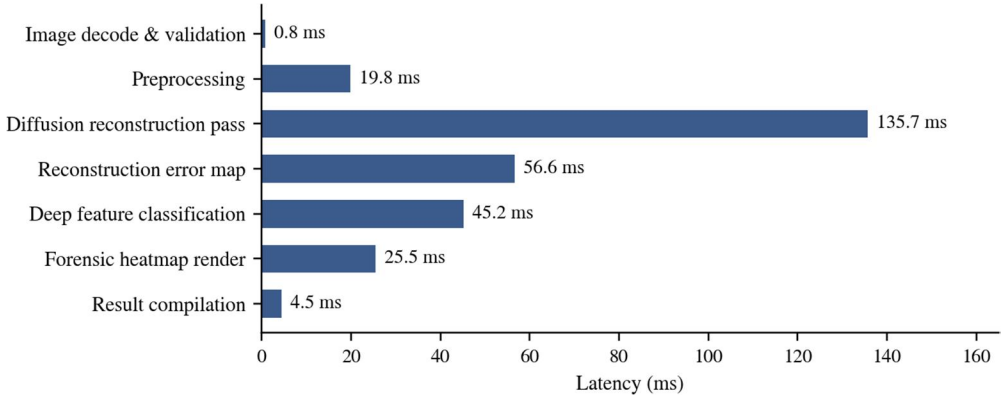


Fig. 4. Latency of each pipeline stage in a single image analysis (total 288.1 ms).

D. Qualitative Output

Figure 5 shows the single-image view: the original image beside its forensic reconstruction map, a verdict ("AI-generated fake", 57.8% synthetic, moderate risk), and a metrics panel with inference latency, resolution and dominant spectral cue. This combination lets a user check *where* the evidence lies rather than rely on the percentage alone.

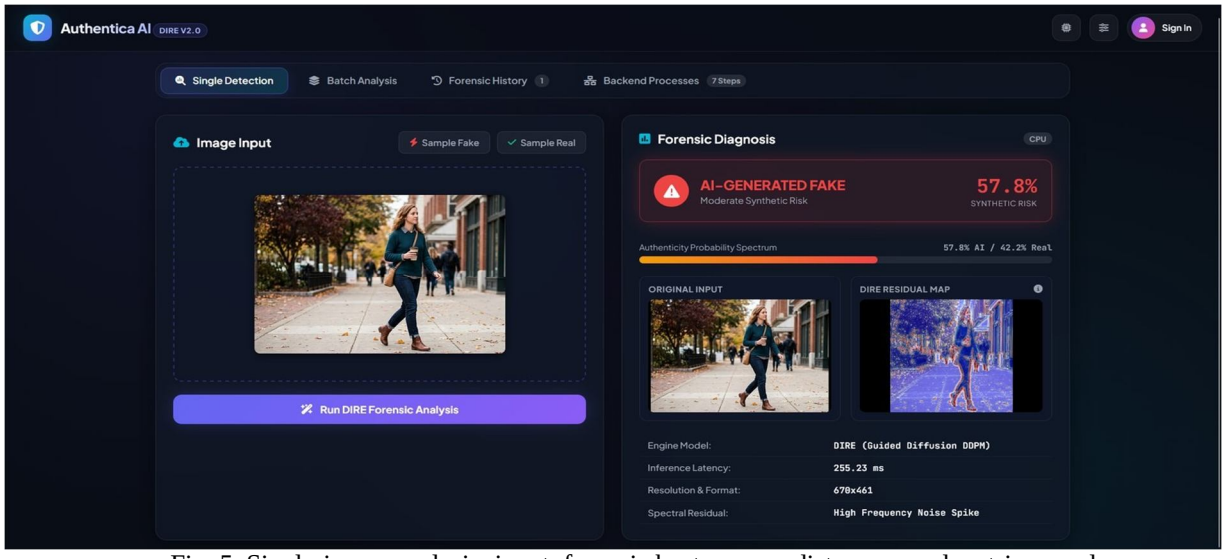


Fig. 5. Single-image analysis: input, forensic heatmap, verdict gauge and metrics panel.

VI. DISCUSSION AND LIMITATIONS

The results above demonstrate that the pipeline runs end to end at interactive speed and that its scores separated the small set of examples tried. They do not establish detection accuracy, and the following limitations should be kept in mind.

1) No benchmark evaluation. The system has not yet been evaluated on a labelled public dataset, so accuracy, AUC, false-positive and false-negative rates, and cross-generator generalisation are unmeasured. The five-image batch is a demonstration, not a statistical test.

2) Training details. The weights, training data and training procedure of the image and audio models are not described here, and the system can fall back to a pretrained model in demonstration mode. Reported scores should be read as outputs of a prototype.

3) Score repeatability. The same file ("Screenshot 2026-09-18 073749.png") scored 57.8% in single-image mode and 62.0% in the batch run. Both are above the decision threshold, but the difference suggests run-to-run variation, possibly from the random noise injected in the diffusion pass; fixing the random seed or averaging several passes is a natural remedy. This is a hypothesis that remains to be tested.

4) Modality coverage. Quantitative results here concern images. Audio detection is implemented but not benchmarked in this paper, and video analysis is not yet implemented.

5) Hardware. Latency was measured on a single CPU configuration and includes a longer first-run time (400.2 ms for the first batch item).

6) Responsible use. A probabilistic score is evidence, not proof. Detector output should not be the sole basis for accusing a person or removing content, and users should be told that false positives and false negatives occur.

VII. CONCLUSION AND FUTURE WORK

Authentica AI shows that a modality-aware router, a hybrid image pipeline that fuses deep features with reconstruction error and frequency-domain forensics, and an explainable heatmap interface can be delivered as a single web platform that runs at interactive speed on a CPU. Its per-stage profile identifies the reconstruction pass as the main cost, and its limitations are stated above.

Future work will: (1) evaluate the system on public labelled benchmarks and report accuracy, AUC and robustness to compression and noise; (2) extend detection to video using a Vision Transformer [6] frame analyser for face-swap, lip-sync and temporal inconsistencies; (3) add Grad-CAM [12] explanations alongside the forensic heatmap; (4) train and fine-tune on newer generators such as Stable Diffusion 3, Midjourney v6 and Sora; (5) add cross-modal consistency analysis so that image, audio and video from the same source can be evaluated jointly; and (6) provide a rate-limited, monitored REST API for production use.

©IJRASET: All Rights are Reserved

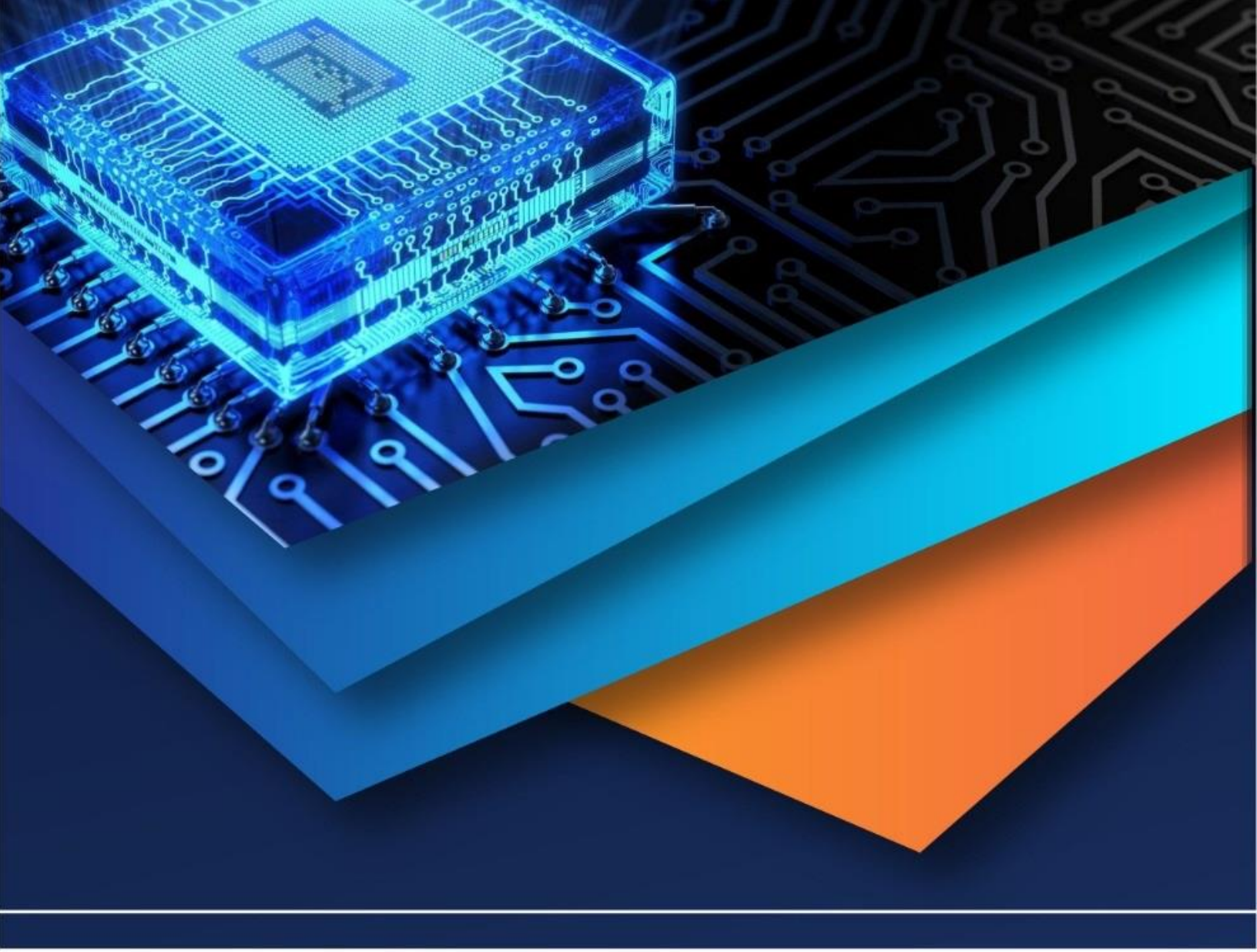
VIII. ACKNOWLEDGMENT

The authors thank Dr. S. Miruna Joe Amali, Head of the Department of Computer Science and Engineering, and the faculty and staff of K.L.N. College of Engineering for their guidance and support during this work.

REFERENCES

[1] M. A. Raza and K. M. Malik, "Multimodaltrace: Deepfake detection using audiovisual representation learning," IEEE Xplore, 2023.
[2] L. Lin, N. Gupta, Y. Zhang, H. Ren, C.-H. Liu, F. Ding, X. Wang, X. Li, L. Verdoliva, and S. Hu, "Detecting multimedia generated by large AI models: A survey," arXiv preprint arXiv:2402.00045, 2024.
[3] K. Lee, Y. Zhang, and Z. Duan, "A multi-stream fusion approach with one-class learning for audio-visual deepfake detection," in Proc. IEEE Int. Workshop on Multimedia Signal Processing (MMSP), 2024, pp. 1–6.
[4] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. 36th Int. Conf. on Machine Learning (ICML), 2019.
[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016.
[6] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Int. Conf. on Learning Representations (ICLR), 2021.
[7] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "DIRE for diffusion-generated image detection," in Proc. IEEE/CVF Int. Conf. on Computer Vision (ICCV), 2023.
[8] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems (NeurIPS), 2020.
[9] I. Goodfellow et al., "Generative adversarial nets," in Advances in Neural Information Processing Systems (NeurIPS), 2014.
[10] R. Durall, M. Keuper, and J. Keuper, "Watch your up-convolution: CNN based generative deep neural networks are failing to reproduce spectral distributions," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2020.
[11] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in Proc. 37th Int. Conf. on Machine Learning (ICML), 2020.
[12] 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. on Computer Vision (ICCV), 2017.
[13] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot… for now," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2020.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)