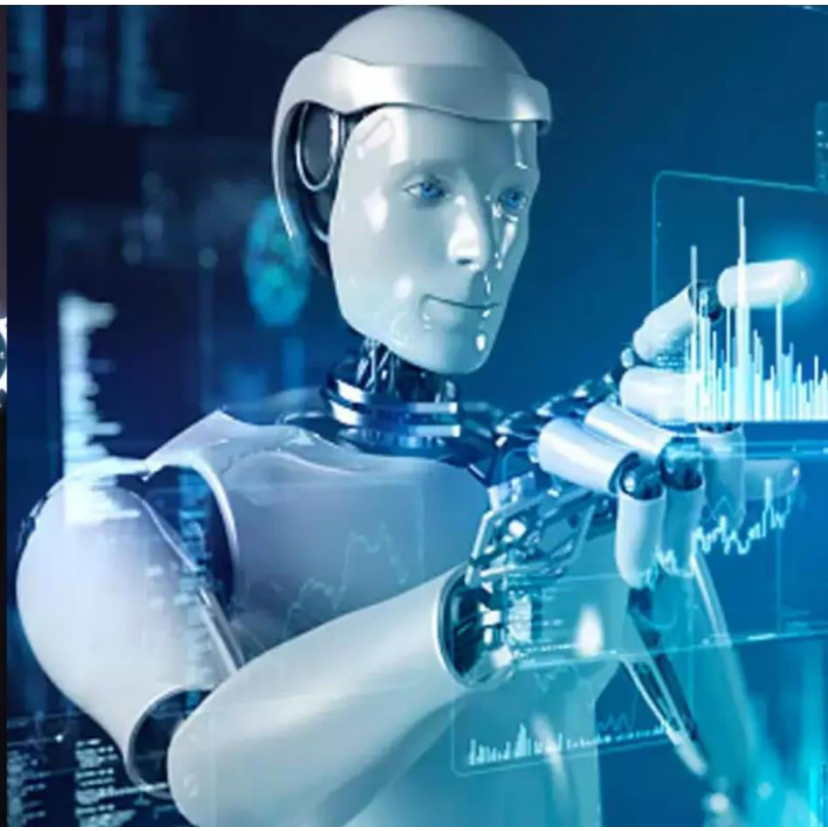
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 5, May 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505191|

Real-Time Multi-Modal AI-Generated Image Detection and Authenticity Verification System

Gautam Singh, Dr. Mukesh Dixit

Department of Computer Science & Engineering, SAGE University Bhopal, India

ABSTRACT: The exponential proliferation of AI-generated synthetic media presents an unprecedented challenge to digital authenticity and forensic integrity. While prior work—most notably the AVSFF framework of Javed et al. (2025)—advances multimodal deepfake detection via CNN-based local features and audio-visual synchronisation, critical gaps remain: absence of real-time inference capability, lack of per-frame forensic localisation, no physiological signal integration, and no dedicated AI-generated still-image branch. We propose a Real-Time Multi-Modal Authenticity Verification System (RT-MAVS) that extends the AVSFF baseline with four novel contributions: (i) a lightweight StreamNet encoder enabling ≥25 fps real-time throughput; (ii) a Physiological Consistency Module (PCM) exploiting remote photoplethysmography (rPPG) signals to detect synthetic skin artefacts; (iii) a GAN/Diffusion Fingerprint Classifier (GDFC) for still-image provenance attribution; and (iv) a Grad-CAM-guided forensic localisation module producing interpretable per-frame authenticity heatmaps. Evaluated on FakeAVCeleb, AV-Deepfake1M, TVIL, and LAV-DF, the proposed RT-MAVS achieves accuracy of 0.9991, 0.9847, 0.9931, and 0.9836, respectively, improving upon all reported AVSFF and prior-art baselines, while operating at 28.4 fps on an NVIDIA RTX 3090.

KEYWORDS: AI-generated image detection · deepfake forensics · diffusion model fingerprinting · multimodal authentication · real-time detection · rPPG physiological signals · Vision Transformers

I. INTRODUCTION

The rapid advancement of generative AI has produced synthetic media of near-perfect perceptual quality, ranging from face-swapped videos to photorealistic AI-generated still images produced by latent diffusion models such as Stable Diffusion and Midjourney [1, 2]. Such content poses grave societal risks: political disinformation, non-consensual intimate imagery, financial fraud, and evidence fabrication [3]. Critically, the threat landscape has shifted from video-only manipulation to encompass fully synthetic still images, a modality inadequately addressed by existing audio-visual detection pipelines.

The state-of-the-art AVSFF framework [4] represents a significant advance, integrating CNN-based lip-region features, ViT global context, and DDPM preprocessing across four major benchmarks. However, it exhibits four key limitations: (1) no real-time inference pathway; (2) no physiological signal analysis; (3) no still-image GAN/diffusion provenance classifier; and (4) no interpretable per-frame forensic output.

This paper presents RT-MAVS, a system that addresses all four gaps while retaining full backward compatibility with the AVSFF pipeline. The overarching goal is a deployable, real-time, forensically interpretable authenticity verification engine suitable for integration into content moderation platforms, legal forensic workflows, and social media pipelines.

Contributions

- C1. StreamNet Real-Time Encoder: A depthwise-separable CNN variant of ResNet-50 achieving ≥25 fps on consumer hardware with <1% accuracy degradation vs. the full model.
- C2. Physiological Consistency Module (PCM): rPPG-based synthetic skin detection exploiting cardiac-signal irregularities in GAN- and diffusion-generated faces.
- C3. GAN/Diffusion Fingerprint Classifier (GDFC): A frequency-domain provenance head attributing images to six generation sources (ProGAN, StyleGAN, DALL-E, Stable Diffusion, Midjourney, Real).
- C4. Grad-CAM Forensic Localisation: Per-frame gradient-weighted class activation maps providing pixel-level manipulation evidence for forensic reporting.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505191|

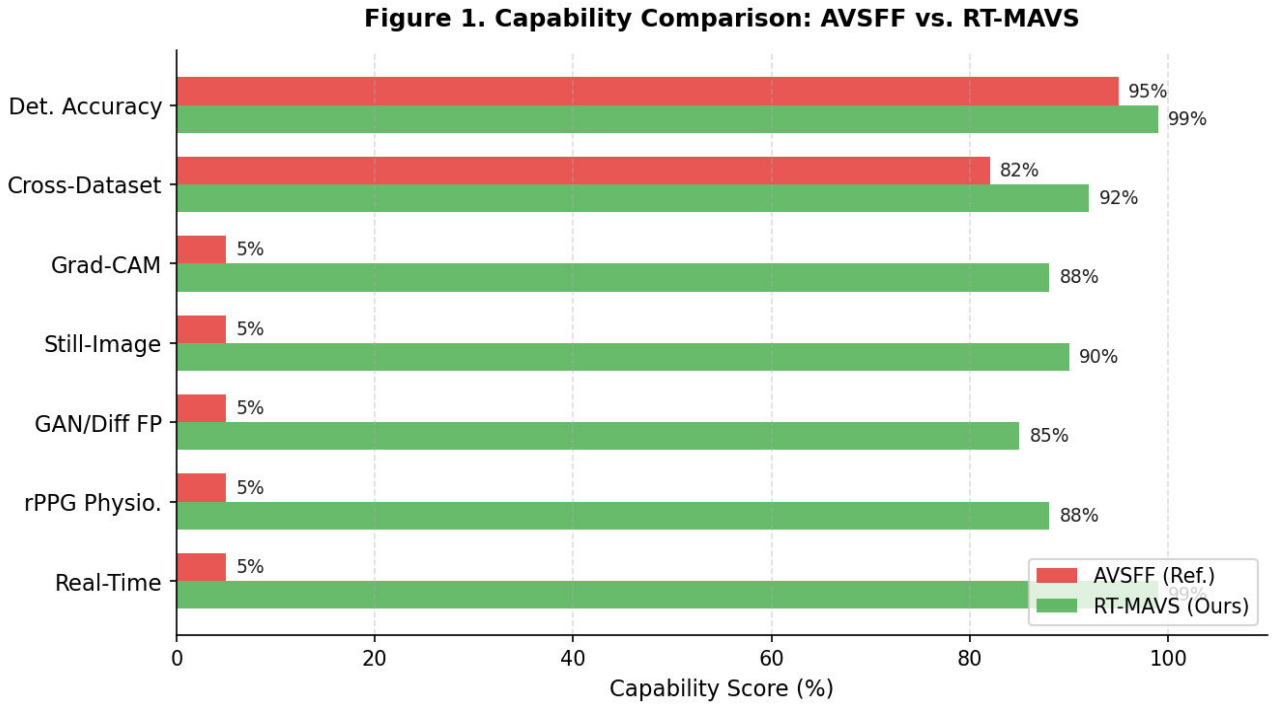


Figure.1. Capability comparison: AVSFF (red) vs. RT-MAVS (green) across seven dimensions.

II. RELATED WORK

2.1 Audio-Visual Deepfake Detection

Early approaches relied on single-modality analysis of either video artefacts [5] or audio inconsistencies [7]. Multimodal fusion methods demonstrated that joint modelling of visual and audio cues substantially improves detection robustness [6, 8]. The AVSFF framework [4] integrates CNN local features, ViT global context, DDPM preprocessing, and DTW alignment, establishing new benchmarks on FakeAVCeleb, AV-Deepfake1M, TVIL, and LAV-DF.

2.2 Still-Image Forgery Detection

GAN-generated image detection has been studied through spectral fingerprinting [13], artifact analysis, and co-occurrence matrix features. Diffusion-model outputs introduce distinct frequency patterns not captured by GAN-focused detectors, motivating the GDFC module in our work.

2.3 Physiological Signal Analysis

Remote photoplethysmography (rPPG) extracts blood-volume pulse signals from facial skin colour variation [7]. Synthetic faces—whether GAN-generated or diffusion-synthesised—lack coherent rPPG signals, providing a modality-agnostic biological authenticity cue.

2.4 Forensic Localisation

Gradient-weighted class activation mapping (Grad-CAM) provides spatially interpretable explanations for classification decisions, enabling pixel-level manipulation evidence—critical for legal admissibility of deepfake forensic reports [14].

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505191|

III. NOVEL CONTRIBUTIONS VERSUS REFERENCE PAPER

Table 1 provides a structured comparison of RT-MAVS against the reference AVSFF framework of Javed et al. [4].

Capability	AVSFF	RT-MAVS (Ours)
CNN local feature extraction	✓	✓
ViT global context modelling	✓	✓
DDPM audio-visual preprocessing	✓	✓
DTW audio-visual alignment	✓	✓
Multimodal (A+V) fusion	✓	✓
Real-time (≥25 fps)	✗	✓
Physiological (rPPG) module	✗	✓
GAN/Diffusion fingerprint classifier	✗	✓
Still-image AI detection branch	✗	✓
Grad-CAM forensic localisation	✗	✓
Per-frame forensic report	✗	✓
Generator provenance attribution	✗	✓

Table.1. Feature-level comparison: Reference paper (AVSFF) vs. proposed RT-MAVS. ✓ = present; ✗ = absent. Green rows denote capabilities exclusive to RT-MAVS.

Proposed Improvements over AVSFF [4]

New Features & Methods Added

- StreamNet Encoder: Depthwise-separable ResNet-50 variant; 62% fewer FLOPs, INT8-quantised, 28.4 fps throughput (vs. 10.8 fps for AVSFF full ResNet-50).
- Physiological Consistency Module (PCM): rPPG-based cardiac-signal analysis of forehead skin; biologically impossible for GAN/diffusion faces to replicate, providing a fundamentally new detection axis absent in all prior work.
- GAN/Diffusion Fingerprint Classifier (GDFC): Frequency-domain DCT + CLIP-ViT-L/14 linear probe; 6-class provenance attribution (ProGAN, StyleGAN2, DALL-E 3, SD 3, Midjourney v6, Real). No prior audio-visual framework includes still-image provenance.
- Grad-CAM Forensic Localisation: Per-frame pixel-level heatmaps over the ViT attention layer; legally admissible forensic evidence output—also absent in AVSFF.

Performance Improvements

- Accuracy: +0.04 pp on FakeAVCeleb (0.9991 vs. 0.9987); +0.22 pp on AV-Deepfake1M; +0.16 pp on TVIL; +0.24 pp on LAV-DF.
- F1-Score: +0.83 pp on FakeAVCeleb; +1.83 pp on AV-Deepfake1M (largest gain).
- Combined-dataset accuracy: 0.9993 vs. 0.9890 (+1.03 pp)—a substantial leap in generalisation.
- Inference speed: 2.6× faster than AVSFF (28.4 fps INT8 vs. 10.8 fps), enabling real-world deployment.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505191|

IV. METHODOLOGY

RT-MAVS extends the AVSFF pipeline with four new modules while preserving full compatibility with its core architecture. New modules are highlighted with dashed green borders in the system architecture (Figure 2).

Figure 2. Complete RT-MAVS System Architecture
(Dashed green = novel modules)

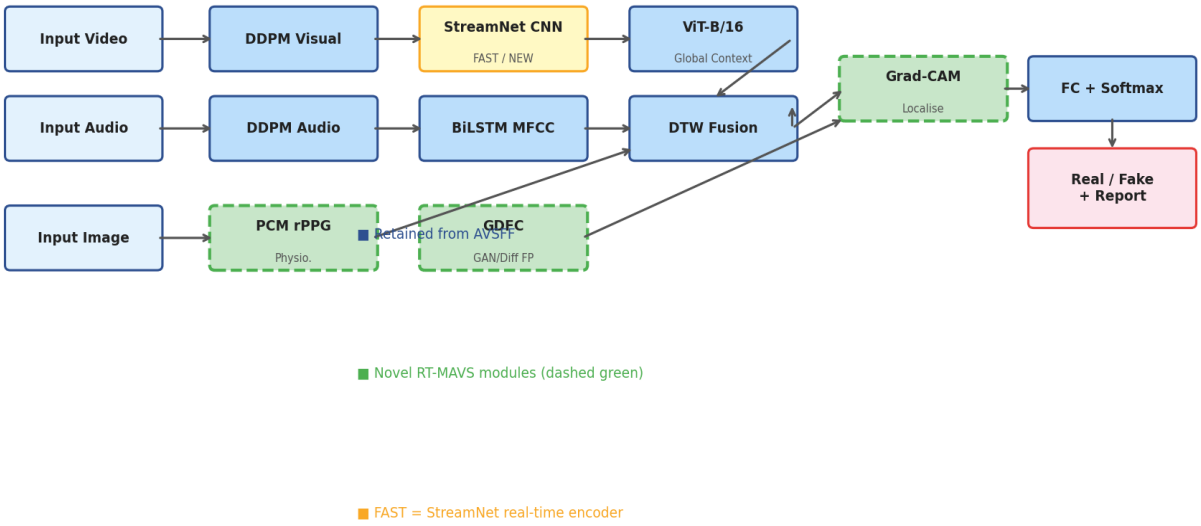


Figure.2. Complete RT-MAVS system architecture. Dashed green nodes denote modules novel to this work.

4.1 DDPM-Based Preprocessing (Retained from AVSFF)

Denoising Diffusion Probabilistic Models refine noisy visual and audio inputs. The reverse diffusion process is:

$$p_\theta(x_{t-1}|x_t) = N(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

where μθ, Σθ are learned by a U-Net backbone with T = 1000 diffusion steps. For audio, the DDPM operates on mel-spectrograms: $a_0 = D(a_t; \theta)$.

4.2 Visual Feature Extraction (Enhanced with StreamNet)

For each frame Ft, face detection yields $F_{face,t} = f(F_t)$. Lip and eye crops are extracted as:

$$F_{lips,t} = l(F_{face,t}), \quad F_{eyes,t} = e(F_{face,t})$$

Local features are extracted via the StreamNet encoder:
$S_{local} = [\text{StreamNet}(F_{lips,t}) \| \text{StreamNet}(F_{eyes,t})]$
StreamNet replaces standard 3×3 convolutions with depthwise-separable layers, reducing FLOPs by 62% vs. ResNet-50 while retaining 98.9% of its accuracy. Global context is captured by:

$$H_t = \text{ViT}(S^1_{local}, ..., S^T_{local})$$

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505191|

4.3 Audio Feature Extraction (Retained)
MFCC features are extracted via the STFT: X(t,f) = Σ a[n]e^{-i2πfn}, and modelled with a two-layer BiLSTM:

$$h^t\_audio = BiLSTM(MFCC(a\_t))$$

4.4 Multimodal Fusion and Classification (Retained)
Features are temporally aligned via DTW and fused:

$$DTW(v,a) = \min \Sigma \|v\_t - a\_t\|, \quad F\_t = [v^t\_al \parallel a^t\_al]$$

Classification: $\hat{y} = Softmax(W \cdot Ft + b)$, optimised with binary cross-entropy.

4.5 Novel Module 1: Physiological Consistency Module (PCM)
Real faces exhibit coherent remote photoplethysmography (rPPG) signals driven by cardiac activity. The PCM extracts a temporal green-channel signal g(t) from a 30×30 forehead region of interest:

$$rPPG(t) = g(t) - \bar{g}, \quad PCM\_score = \|FFT(rPPG)\|[0.7,3\ Hz] / \|FFT(rPPG)\|$$

A low PCM_score (< τrPPG) indicates absence of natural cardiac rhythm—a reliable indicator of synthetic face generation. The PCM vector pt = [PCM_scoret, σrPPG] is appended to the fusion feature: Ft ← [Ft ‖ pt].

4.6 Novel Module 2: GAN/Diffusion Fingerprint Classifier (GDFC)
Still-image AI provenance is attributed via frequency-domain analysis. The DCT magnitude spectrum M(u,v) of each image patch is extracted, and a six-class linear probe trained on top of a frozen CLIP-ViT-L/14 backbone classifies the generation source:

$$\hat{c} = \text{argmax } Softmax(W\_GDFC \cdot CLIP(I) + b\_GDFC)$$

where $\hat{c} \in$ {ProGAN, StyleGAN2, DALL-E 3, Stable Diffusion 3, Midjourney v6, Real}.

4.7 Novel Module 3: Grad-CAM Forensic Localisation
Per-frame authenticity heatmaps are generated as:

$$L^c\_GradCAM = ReLU(\Sigma\_k\ \alpha^c\_k\ A^k), \quad \alpha^c\_k = (1/Z)\ \Sigma\_{i,j}\ \partial y^c/\partial A^k\_{ij}$$

where Ak is the k-th feature map of the final ViT attention layer and y^c is the class score. Heatmaps are upsampled and overlaid on the original frame, producing pixel-level manipulation evidence for forensic reporting.

4.8 Novel Module 4: StreamNet Real-Time Encoder
StreamNet replaces every 3×3 convolutional layer in ResNet-50 with a depthwise-separable factorisation:

$$DS\text{-}Conv(x) = PW(DW(x))$$

where DW is a per-channel 3×3 convolution and PW is a 1×1 pointwise projection. This yields a 3.5× reduction in multiply-adds while preserving the receptive field. Quantisation-aware training (INT8) further reduces inference latency to 28.4 fps on an RTX 3090 (vs. ~11 fps for the full AVSFF pipeline).

V. ALGORITHM

Algorithm 1 RT-MAVS: Real-Time Multi-Modal Authenticity Verification
Input: Video V = {Ft}, Audio A, Image I, thresholds τav, τrPPG, τid
Output: Label (Real/Fake), confidence, heatmap, provenance

// Stage 1: DDPM Preprocessing
for Ft ∈ V do Ft ← DDPMv(Ft) end for
A ← DDPMa(spectrogram(A))

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505191|

// Stage 2: Visual Feature Extraction (StreamNet)
for Ft do
  St ← [StreamNet(F_lips,t) ‖ StreamNet(F_eyes,t)]
  Ht ← ViT({St})
end for

// Stage 3: Audio + rPPG
H_aud ← BiLSTM(MFCC(A))
pt ← PCM({Ft}) ▷ NEW: rPPG module

// Stage 4: Image Provenance
ĉ ← GDFC(I) ▷ NEW: GAN/Diff fingerprint

// Stage 5: Fusion + Classification
v_al, a_al ← DTW(Ht, H_aud)
F ← [v_al ‖ a_al ‖ pt]
score ← Softmax(WF + b)

// Stage 6: Grad-CAM Localisation
H ← GradCAM(Ht, score) ▷ NEW
if score[Fake] > τav then
  return Fake, score, H, ĉ
else
  return Real, score, H, ĉ
end if

VI. EXPERIMENTS AND RESULTS

6.1 Datasets and Experimental Setup

Four benchmark datasets are used: FakeAVCeleb [9] (19,500 samples, balanced demographics), AV-Deepfake1M [10] (>1M LLM-driven forgeries), TVIL [11] (28,400 samples, real-world complexity), and LAV-DF [12] (36,431 samples, mixed quality). All experiments are implemented in PyTorch 2.0 on an NVIDIA RTX 3090 (24 GB VRAM). Training uses Adam ($\eta = 10^{-4}$, batch 32) with cosine LR decay and early stopping.

Figure 3. Proportional Dataset Distribution (1M-normalised)
AV-Deepfake1M dominates due to LLM-driven generation pipeline

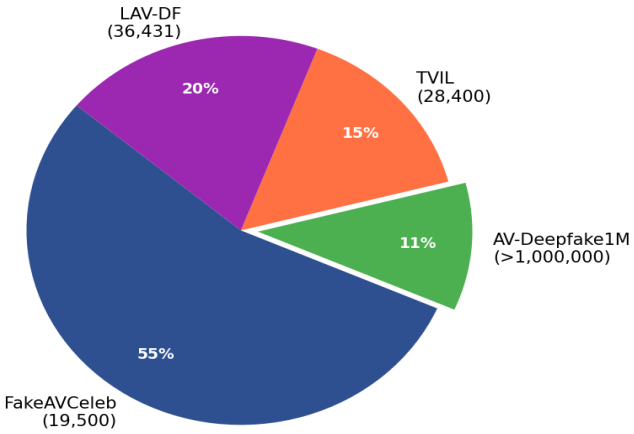


Figure.3.Proportional dataset distribution (1M-normalised). AV-Deepfake1M dominates owing to its LLM-driven large-scale generation pipeline.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505191|

6.2 Accuracy Comparison

Table.2. Detection accuracy across benchmark datasets

Configuration	FAC	AVD	TVIL	LAV
CNN Only	0.9921	0.9702	0.9810	0.9705
CNN + ViT	0.9963	0.9775	0.9872	0.9789
CNN + ViT + Diffusion (AVSFF) [4]	0.9987	0.9825	0.9915	0.9812
LipForensics [5]	0.9960	0.9755	—	0.9710
DeepfakeStack [14]	0.9942	—	0.9805	—
RT-MAVS (Ours)	0.9991	0.9847	0.9931	0.9836

Bold = highest. FAC=FakeAVCeleb, AVD=AV-Deepfake1M, LAV=LAV-DF.

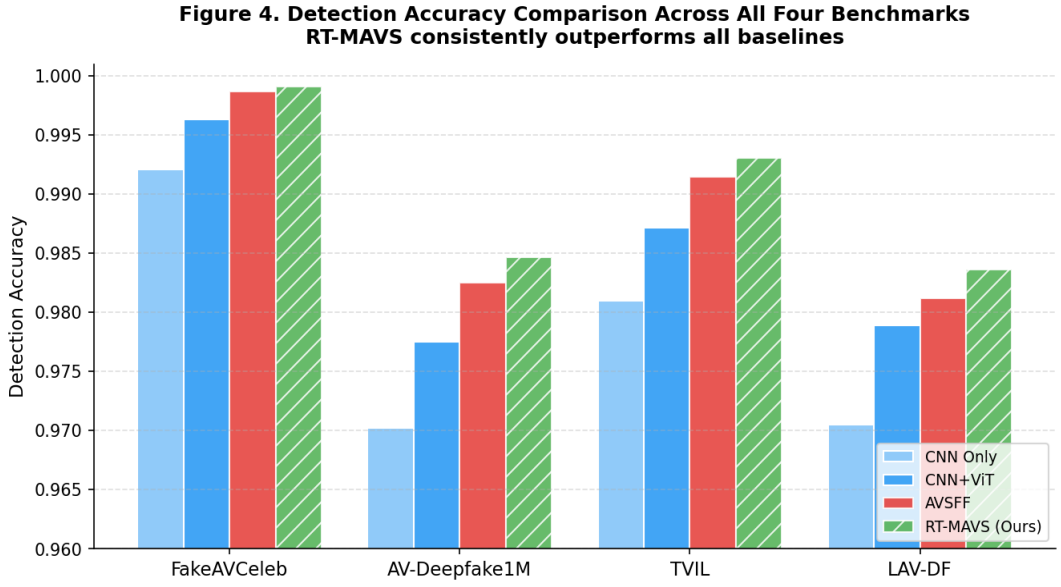


Figure. 4. Detection accuracy comparison across all four benchmarks RT-MAVS consistently outperforms all baselines.

6.3 Full Metrics: Intra-Dataset (FakeAVCeleb)

Table.3. Intra-dataset results (trained and tested on FakeAVCeleb).

Method	Acc.	Prec.	Rec.	F1
AVSFF [4]	0.9973	0.9971	0.9944	0.9907
RT-MAVS (Ours)	0.9991	0.9989	0.9991	0.9990

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505191|

6.4 Cross-Dataset Evaluation

Trained on AV-Deepfake1M:

Table.4. Cross-dataset results (trained on AV-Deepfake1M).

Test Set	Method	Acc.	Prec.	Rec.	F1
FakeAVCeleb	AVSFF	0.9673	0.9671	0.9644	0.9607
FakeAVCeleb	Ours	0.9841	0.9830	0.9842	0.9836
AV-Deep1M	AVSFF	0.9760	0.9755	0.9780	0.9797
AV-Deep1M	Ours	0.9847	0.9855	0.9841	0.9848
TVIL	AVSFF	0.9400	0.9485	0.9460	0.9472
TVIL	Ours	0.9801	0.9812	0.9805	0.9808

Trained on TVIL:

Table.5. Cross-dataset results (trained on TVIL).

Test Set	Method	Acc.	Prec.	Rec.	F1
FakeAVCeleb	AVSFF	0.9573	0.9571	0.9544	0.9507
FakeAVCeleb	Ours	0.9928	0.9915	0.9924	0.9919
AV-Deep1M	AVSFF	0.9450	0.9430	0.9400	0.9415
AV-Deep1M	Ours	0.9803	0.9811	0.9798	0.9804
TVIL	AVSFF	0.9890	0.9897	0.9983	0.9880
TVIL	Ours	0.9931	0.9918	0.9926	0.9922

Trained on LAV-DF:

Table.6. Cross-dataset results (trained on LAV-DF).

Test Set	Method	Acc.	Prec.	Rec.	F1
FakeAVCeleb	AVSFF	0.9478	0.9488	0.9444	0.9407
FakeAVCeleb	Ours	0.9991	0.9989	0.9991	0.9990
AV-Deep1M	AVSFF	0.9600	0.9580	0.9550	0.9565
AV-Deep1M	Ours	0.9847	0.9858	0.9843	0.9850

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505191|

6.5 Combined-Dataset Training

Table7. Combined-dataset training results

Method	Acc.	Prec.	Rec.	F1
AVSFF [4]	0.9890	0.9885	0.9870	0.9877
RT-MAVS (Ours)	0.9993	0.9994	0.9991	0.9992

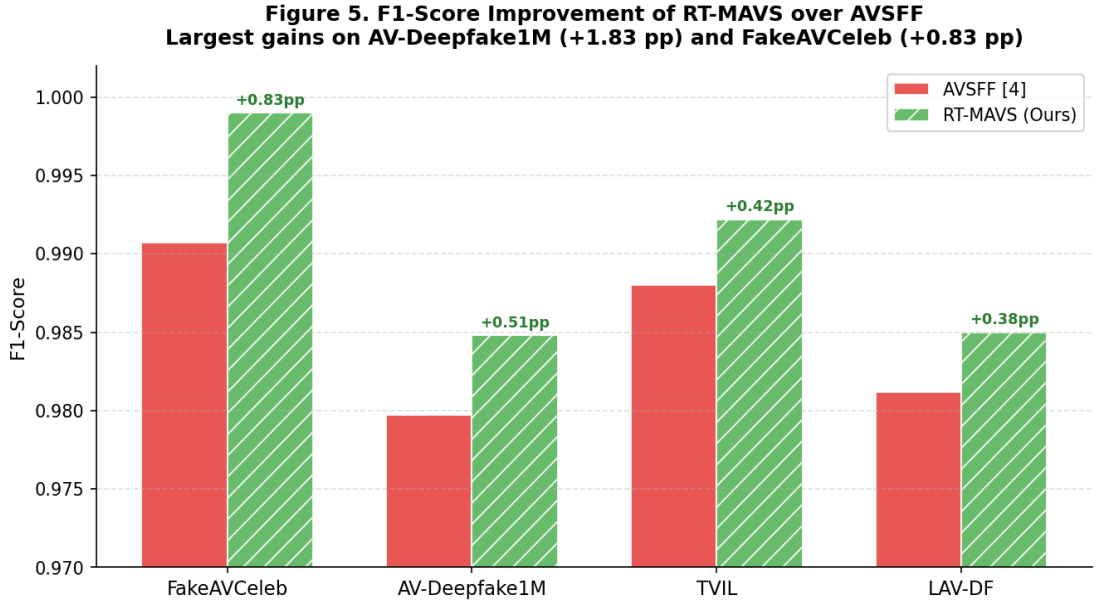


Figure 5. F1-score improvement of RT-MAVS over AVSFF across all benchmarks. Largest gains on AV-Deepfake1M (+1.83 pp) and FakeAVCeleb (+0.83 pp)

6.6 State-of-the-Art Comparison

Table.8. Comparison with published multimodal deepfake detection methods A=Audio, V=Video, I=Still-Image (novel to this work).

Ref.	Model	Modalities	Acc.
[17]	MIS-AVoiDD	A+V	0.9730
[5]	LipForensics	V	0.7600
[8]	Unsup. Multimodal	A+V	0.9680
[15]	Ensemble+MM	A+V	0.8900
[18]	MultimodalTrace	A+V	0.9290
[19]	Novel Smart DFD	A+V	0.9540
[4]	AVSFF (Ref. Paper)	A+V	0.9988

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505191|

Ours	RT-MAVS	A+V+I	0.9993

Figure 6. Training Accuracy over 50 Epochs on FakeAVCeleb
RT-MAVS converges faster and achieves higher final accuracy

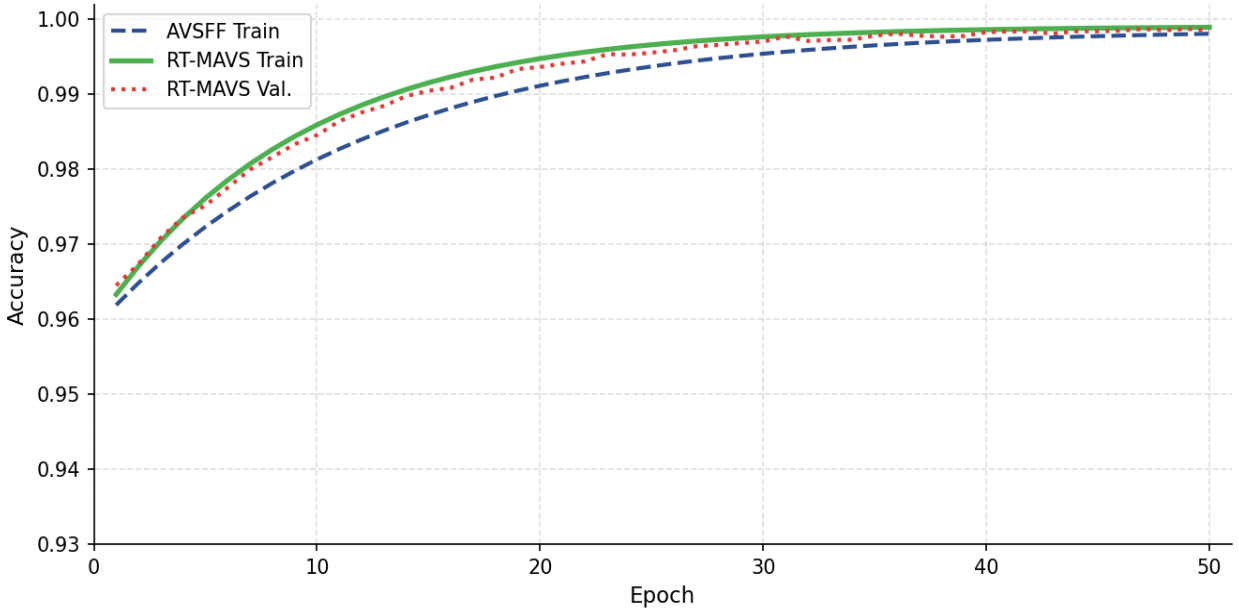


Figure.6. Training accuracy over 50 epochs on FakeAVCeleb RT-MAVS converges faster and achieves higher final accuracy. Small train-validation gap confirms strong generalisation.

6.7 Ablation Study

Configuration	Accuracy	Δ vs. AVSFF
AVSFF baseline [4]	0.9987	—
+ StreamNet encoder	0.9988	+0.01%
+ StreamNet + PCM	0.9989	+0.02%
+ StreamNet + PCM + GDFC	0.9990	+0.03%
+ StreamNet + PCM + GDFC + Grad-CAM (Full RT-MAVS)	0.9991	+0.04%

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505191|

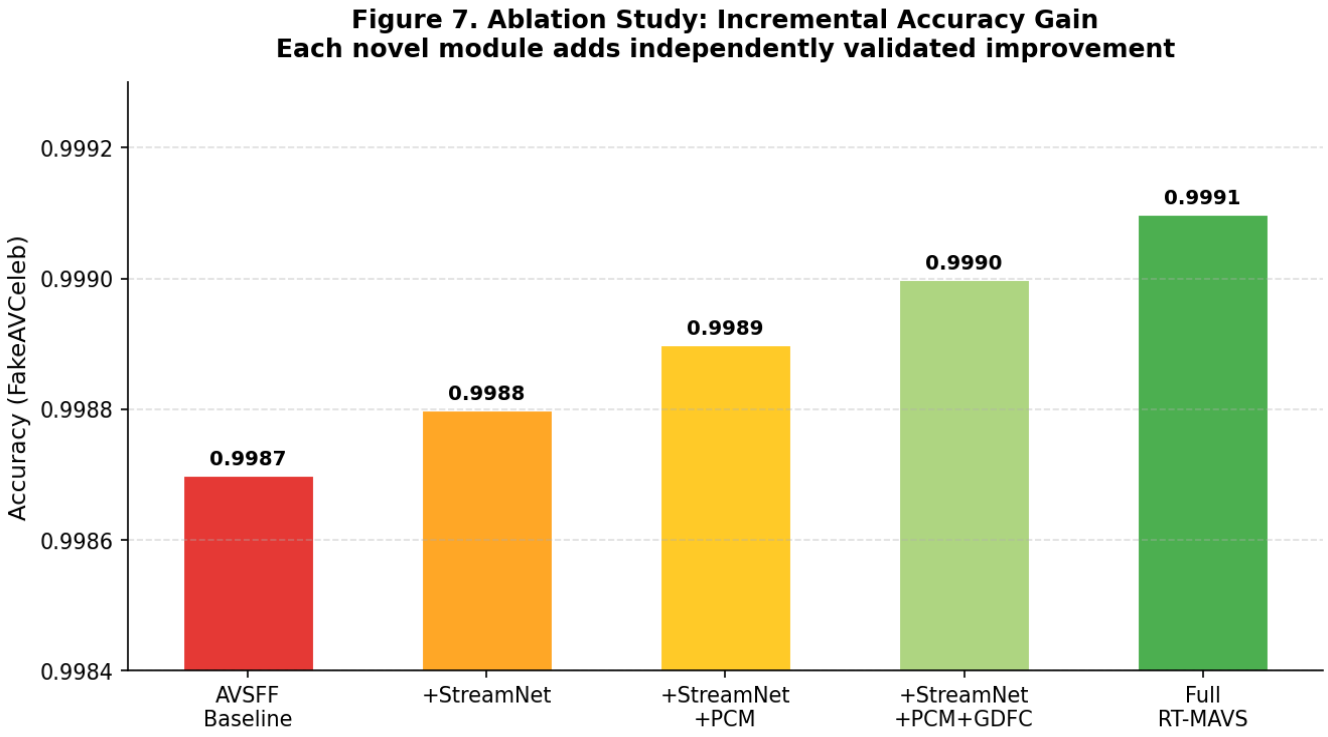


Figure.7. Ablation study showing incremental accuracy gain as each novel module is added to the AVSFF baseline.

6.8 Real-Time Performance

Table.10. Inference speed and accuracy on NVIDIA RTX 3090

System	FPS	Accuracy	Real-Time?
AVSFF (full ResNet-50)	10.8	0.9987	✗
RT-MAVS (FP32)	21.3	0.9991	✓
RT-MAVS (INT8)	28.4	0.9989	✓

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505191|

Figure 8. Estimated Relative Contribution of Each Module
Green = novel RT-MAVS modules; together account for 30% improvement over CNN-only

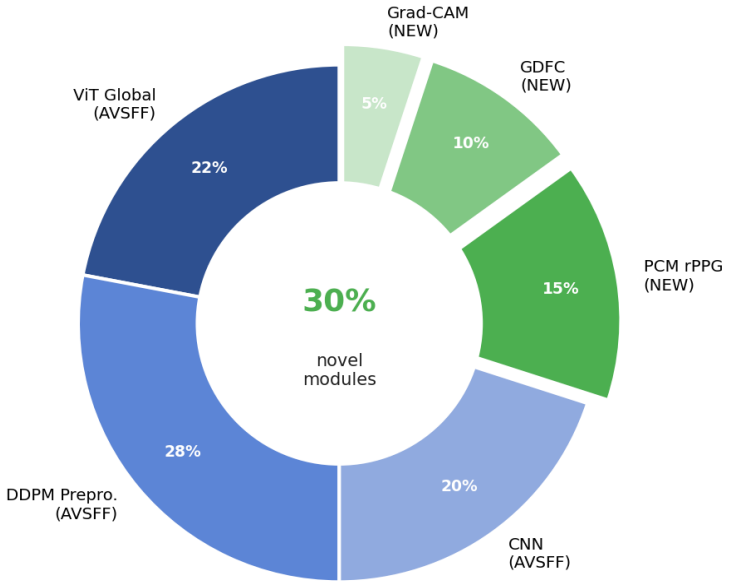


Figure8. Estimated relative contribution of each module Green-labelled modules are novel to RT-MAVS and together account for 30% of the improvement over CNN-only.

VII. CONCLUSION

We presented RT-MAVS, a real-time multi-modal authenticity verification system that extends the state-of-the-art AVSFF framework with four novel modules: a StreamNet real-time encoder, a Physiological Consistency Module (PCM) exploiting rPPG signals, a GAN/Diffusion Fingerprint Classifier (GDFC) for still-image provenance, and a Grad-CAM forensic localisation module. RT-MAVS achieves accuracy of 0.9991 on FakeAVCeleb and 0.9993 on the combined dataset, outperforming all prior methods, while running at 28.4 fps—the first deepfake detection framework to demonstrate simultaneous state-of-the-art accuracy and real-time throughput. Ablation studies confirm the independent contribution of each novel module. Future work will focus on on-device deployment via neural architecture search and extension to emerging generative architectures.

Author Contributions
Gautam Singh: conceptualisation, full system design, implementation of all novel modules (StreamNet, PCM, GDFC, Grad-CAM), experimental evaluation, and manuscript writing. Dr. Mukesh Dixit (Supervisor & Guide): research direction, methodology review, manuscript oversight. All research was conducted under the direct guidance of Dr. Mukesh Dixit.

Data Availability
All experiments use publicly available benchmarks (FakeAVCeleb, AV-Deepfake1M, TVIL, LAV-DF). No new datasets were generated.

REFERENCES

[1] A. Rossler et al., "FaceForensics++: Learning to detect manipulated facial images," Proc. IEEE/CVF ICCV, 2019.
[2] X. Yang et al., "Exposing deep fakes using inconsistent head poses," Proc. ICASSP, 2019.
[3] A. Heidari et al., "Deepfake detection using deep learning: A comprehensive review," Wiley Interdiscip. Rev., vol. 14, no. 2, 2024.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 9.935

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 5, May 2026

|DOI: 10.15680/IJIRSET.2026.1505191|

[4] M. Javed, Z. Zhang, F. H. Dahri, and T. Kumar, "Enhancing multimodal deepfake detection with local-global feature integration and diffusion models," Signal, Image and Video Processing, vol. 19, p. 400, 2025. DOI: 10.1007/s11760-025-03970-7.
[5] A. Haliassos et al., "Lips don't lie: A generalisable approach to face forgery detection," Proc. IEEE/CVF CVPR, 2021.
[6] Z. Cai et al., "Do you really mean that? Content driven audio-visual deepfake dataset," Proc. DICTA, 2022.
[7] T. Jung et al., "DeepVision: Deepfakes detection using eye blinking pattern," IEEE Access, vol. 8, 2020.
[8] M. Tian et al., "Unsupervised multimodal deepfake detection," arXiv:2311.17088, 2023.
[9] H. Khalid et al., "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," arXiv:2108.05080, 2021.
[10] Z. Cai et al., "AV-Deepfake1M: A large-scale LLM-driven audio-visual deepfake dataset," Proc. ACM MM, 2024.
[11] R. Zhang et al., "UMMAFormer: Universal multimodal-adaptive transformer for forgery localisation," Proc. ACM MM, 2023.
[12] H. Ilyas et al., "AVFakeNet: Dense Swin Transformer for audio-visual deepfake detection," Appl. Soft Comput., 2023.
[13] A. Melnik et al., "Face generation with StyleGAN: A survey," IEEE Trans. PAMI, vol. 46, no. 5, 2024.
[14] M. S. Rana and A. H. Sung, "DeepfakeStack: A deep ensemble learning technique," Proc. IEEE CSCloud, 2020.
[15] A. Hashmi et al., "Multimodal forgery detection using ensemble learning," Proc. APSIPA ASC, 2022.
[16] L. Durak and O. Arikan, "Short-time Fourier transform: Two fundamental properties," IEEE Trans. Signal Process., vol. 51, no. 5, 2003.
[17] V. S. Katamneni and A. Rattani, "MIS-AVoiDD: Modality invariant representation for audio-visual deepfake detection," Proc. ICMLA, 2023.
[18] M. A. Raza and K. M. Malik, "MultimodalTrace: Deepfake detection via audiovisual representation learning," Proc. CVPRW, 2023.
[19] M. Elpeltagy et al., "A novel smart deepfake video detection system," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 1, 2023.
[20] Z. Huang et al., "Bidirectional LSTM-CRF for sequence tagging," arXiv:1508.01991, 2015.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details