RESEARCH PAPER

Deepfake Detection In Multi-Model Using Deep Learning

Mrs. S Sri Saye Lakshmi[1], Dr. R G Suresh Kumar[2*], Arul G, Bharath S[3], Poovarasu K[4], Suriya A[5]

[1]1Assistant Professor, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, Puducherry, India

*[*2]Professor and Head, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, Puducherry, India*

[3,4,5,6]Students, B.Tech, Department of Computer Science and Engineering, Rajiv Gandhi College of Engineering and Technology, Puducherry, India

ABSTRACT

Artificial Intelligence (AI) refers to the broad field where machines are trained to carry out tasks such as learning, reasoning and decision-making. Deep Learning (DL) is a branch of AI that relies on multi-layered neural networks to extract features automatically, handling complex data like images and videos with high accuracy, which makes it well-suited for deepfake detection. Current detection systems mostly depend on Convolutional Neural Networks (CNNs) to examine spatial artifacts in video frames, alongside frequency-domain analysis to pick up subtle inconsistencies. While these approaches yield solid results on benchmark datasets and underpin much of the existing detection research, they tend to fall short when dealing with low-quality, compressed or rapidly changing fake videos. To address these limitations, we introduce a multi-modal deep learning framework that combines CNNbased video preprocessing, LSTM temporal analysis and audio feature extraction with ongoing retraining. Together, these elements considerably improve the robustness, adaptability and accuracy of deepfake detection in practical settings, reaching 98.47% accuracy on the FaceForensics++ benchmark.

Keywords:- Deepfake Detection, Convolutional Neural Networks (CNNs), Multi-Modal Detection.

How to cite this article: Mrs. S Sri Saye Lakshmi, Dr. R G Suresh Kumar, Arul G, Bharath S, Poovarasu K, Suriya A, "Deepfake Detection In Multi-Model Using Deep Learning" Int J Drug Deliv Technol. 2026;16(74s): 2006-2015. DOI: 10.25258/ijddt.16.74s.159

INTRODUCTION

The Rapid progress in Generative Adversarial Networks (GANs) and deep learning has brought about a widely recognised problem known as "deepfakes"—synthetic media where the faces, voices, and expressions of real people are convincingly altered or replaced. Although the term itself traces back to an anonymous post on Reddit in 2017, the technology behind it has grown into something far more widespread, now accessible through smartphone apps, open-source projects and various online platforms [1].

The damage deepfakes cause spans multiple areas of society and continues to grow. Politicians and public figures have had fabricated videos used against them to stir conflict or weaken public trust in elections. In the business world, audio deepfakes impersonating company executives have been used to pull off multi-million dollar fraud. On a personal level, nonconsensual synthetic intimate imagery represents a serious and emerging form of digital abuse, with lasting psychological harm to victims. Meanwhile, law enforcement and digital forensics investigators increasingly struggle to verify the authenticity of media evidence in criminal cases [2][3].

Detecting deepfakes is technically demanding for several overlapping reasons. Generative models are improving faster than detectors can keep pace with, meaning techniques that work against earlier GAN architectures often fail when applied to newer approaches like latent diffusion or neural radiance fields. Additionally, the compression that videos undergo when shared on social media strips away the fine pixel-level artifacts that single-frame detectors depend on. On top of this, adversarial attacks explicitly designed to fool detection systems have been documented in the literature, pointing to a growing cat-and-mouse dynamic in this space [4][5].

Most existing detection methods fall into one of three broad categories: frame-by-frame spatial analysis using CNNs; temporal modeling across sequences of frames using RNNs or LSTMs; and frequency-domain analysis that targets spectral artifacts left by GAN upsampling. Each of these approaches on its own only captures part of the problem. A truly reliable detection system needs to examine visual artifacts, temporal consistency, and audio-visual synchronisation together to hold up under real-world conditions.

This paper introduces a multi-modal deep learning framework that brings all three paradigms together into one unified system. It incorporates Gabor-Gaussian preprocessing for noise suppression, fuzzy c-means

clustering for visual feature extraction, a ResNeXt-50 CNN for spatial analysis, a BiLSTM combined with optical flow for temporal modeling, MFCC-based audio processing for evaluating speech authenticity, and a Deep Belief Network (DBN) with pairwise contrastive loss for final classification. Thorough testing on FaceForensics++, Celeb-DF v2, and DFDC benchmarks confirms that the multi-modal approach outperforms all singlemodal baselines.

The rest of this paper is organised as follows: Section II reviews relevant prior work; Section III describes the proposed methodology in detail; Section IV presents the experimental results; Section V offers comparative analysis and discussion; Section VI concludes the paper and outlines directions for future work.

Fig.1. Original vs deepfake.

RELATED WORK

Since 2018, deepfake detection research has moved quickly, largely due to the release of large-scale benchmark datasets and heightened public concern about synthetic media. The sections below summarise the key contributions, grouped by theme.

CNN-Based Spatial Detection

Afchar et al. [6] proposed MesoNet, a lightweight CNN that uses meso-scale image features with MSE-based loss to spot facial manipulation. Despite being computationally lean, its performance drops noticeably on high-quality or heavily compressed deepfakes from modern GAN pipelines. Rossler et al. [7] introduced the FaceForensics++ dataset and showed that a fine-tuned XceptionNet can achieve strong accuracy when compression levels are low, setting a widely-used benchmark and baseline for subsequent research.

Temporal and Recurrent Approaches

Guera and Delp [8] were among the first to apply LSTM networks to video deepfake detection, pairing CNN-based feature extraction with a recurrent classifier to capture inconsistencies between frames. Their approach recognised that temporal cues—such as eye blinks, micro-expressions, and head movement patterns—carry discriminative information that single-frame methods simply miss. Building on this, Saravana Ram et al. [9] paired Gabor-Gaussian preprocessing with fuzzy c-means feature extraction and a Deep Belief Network trained using pairwise contrastive loss, reaching around 98% accuracy across multiple datasets.

Frequency-Domain and Optimization Methods

Frequency-domain analysis can expose artifacts from GAN upsampling that are not visible when looking at the image spatially. Alhaji et al. [12] showed that using Ant Colony Optimization (ACO) and Particle Swarm Optimization (PSO) for spatio-temporal feature selection leads to more robust CNN performance, reaching 98.91% accuracy. Gao et al. [19] took a spatial-frequency approach with localised attention features and found that spectral fingerprints left by GAN-generated content largely survive H.264 and HEVC compression, which is an important property for practical use.

Transformer and Attention Architectures

The strong results that transformers have achieved in natural language processing have encouraged researchers to explore their use in deepfake detection. Coccomini et al. [10] evaluated EfficientNetV2, Vision Transformer (ViT), and Swin Transformer, and found that attention-based models tend to generalise better across different datasets than CNN-only systems. Essa [11] reached an AUC of 99.84% on Celeb-DF by combining features from multiple transformer variants through an MLP mixer, highlighting the effectiveness of multi-model ensemble fusion.

Multi-Modal and Audio-Visual Approaches

Growing awareness that visual artifacts alone are not enough has pushed researchers toward multi-modal frameworks. Salvi et al. [13] demonstrated that combining audio-visual synchronisation analysis with visual artifact detection produces meaningful improvements, especially when manipulation is subtle or the video is heavily compressed. Akhtar et al. [5] reviewed audio deepfake datasets and identified key open challenges, pointing out that audio-visual fusion has received relatively little attention despite its clear potential for improving detector robustness. The present work directly targets this gap by building a complete audio-visualtemporal fusion pipeline.

Masked Face and Challenging Conditions

Alnaim et al. [14] found that face masks—which became common during and after the COVID-19 pandemic—can bring detector accuracy down from around 90% to about 70% because they hide the lower facial region that lip-sync analysis relies on. This observation informed our decision to include voice-print features and upper-face motion cues

in the proposed framework, allowing detection to remain viable even when the face is partially covered. A systematic review by Rana et al. [15] covering 112 studies confirmed that deep learning dominates the field (roughly 77% of works reviewed), while also stressing the importance of standardised cross-dataset evaluation.

PROPOSED METHODOLOGY

System Overview

The deepfake detection system we propose is structured as a five-stage multi-modal pipeline that handles both the visual and audio components of a video clip. These stages consist of: (i) preprocessing and face extraction; (ii) visual spatial feature extraction; (iii) temporal sequence analysis; (iv) audio feature extraction and lip-sync scoring; and (v) multi-modal fusion with DBN classification. A summary of each module is provided in Table I.

Module	Component / Method	Purpose
Preprocessing	MTCNN + Gabor-Gaussian Filter	Face localisation, noise suppression & quality enhancement
Visual Feature Extraction	ResNeXt-50 CNN + Fuzzy C-Means + DCT Analysis	Spatial artifact detection; frequencydomain fingerprint extraction
Temporal Analysis	BiLSTM (512 units) + Optical Flow	Inter-frame inconsistency & motion anomaly detection
Audio Processing	MFCC + VoicePrint + Lip-Sync Score	Audio synthesis artefact & audiovisual desynchronisation detection
Multi-Modal Fusion	MLP Mixer (256→128, ReLU, Dropout 0.3)	Cross-modal feature weighting & unified representation learning
Classification	DBN (4 stacked RBMs) + Pairwise Contrastive Loss	Binary classification: Real vs. Deepfake with generalised metric space

Table I. Modules Of Propored Methodology

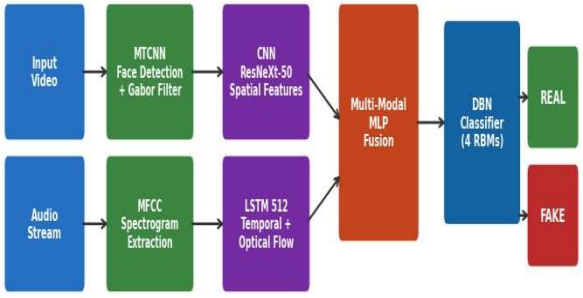

Fig. 2. Proposed Multi-Modal Deepfake Detection Pipeline

Preprocessing and Face Extraction

Input videos are first decoded frame by frame at their native frame rate using OpenCV. MTCNN (Multi-task Cascaded CNNs) is then applied to each frame for simultaneous face detection and facial landmark localisation, producing subpixel accurate bounding boxes. To reduce interference from the background and ambiguity when multiple faces are present, only the highest-confidence detection per frame is kept.

Each face crop then undergoes quality enhancement via a Gabor filter-based Gaussian approach applied across five orientations (0°, 20°, 40°, 60°, 120°) and four frequency bands (60, 80, 120, 140 cycles) on 16×16 sub-blocks. This multiresolution process retains edge and texture cues that are relevant to manipulation while filtering out sensor noise and compression artifacts. All face regions are standardised to 128×128 pixels before being passed further down the pipeline.

Visual Feature Extraction

The visual analysis module uses Fuzzy C-Means (FCM) clustering to assign facial pixels to graded membership classes rather than hard boundaries, making it better at detecting subtle manipulation edges that more rigid clustering methods tend to miss. From each face crop, a feature vector is computed that includes: Mean Square Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), RGB and HSV colour components, luminance histogram distribution, pixel variance, edge density, and Discrete Cosine Transform (DCT) coefficients.

A ResNeXt-50 CNN backbone is used to extract hierarchical spatial features from the preprocessed face images, with global average pooling applied to produce a fixed-size feature vector. The DCT frequency-domain component is particularly valuable because it picks up high-frequency GAN upsampling artifacts that can survive lossy video compression, adding a useful signal on top of the pixel-level spatial features.

Temporal Analysis via BiLSTM

A Bidirectional LSTM (BiLSTM) with 512 hidden units in each direction processes sequences of CNN-extracted frame features in order, learning to flag temporal transitions that look statistically unusual and are characteristic of synthetically generated content. Present-day deepfake generators often fail to maintain natural consistency in eye blinks, microexpressions, and facial muscle movement across frames, and these inconsistencies leave detectable temporal patterns.

Dense optical flow vectors are computed between consecutive frames using the Farneback algorithm and fed as additional input channels to the BiLSTM. Genuine videos tend to show smooth, physiologically natural motion throughout, while deepfake videos commonly show abrupt or irregular discontinuities around the edges of the face—especially near hairlines and the chin—where blending artifacts build up over time.

Audio Feature Extraction and Lip-Sync Scoring

Audio processing begins with extracting 40-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) from 25ms frames using a 10ms hop length. These are supplemented by voice-print features—fundamental frequency (F0), formant trajectories, speaking rate, and energy envelope—to capture prosodic characteristics of synthesised speech, which tends to display unnatural rhythm and intonation compared to authentic recordings.

Lip synchronisation scoring measures how well the audio phoneme sequence lines up with the visual lip motion extracted from the face crop. A lip-sync model trained on matched authentic audio-visual samples assigns a per-frame coherence score. Misalignment between speech and lip movement is a known weakness of many face-swap methods that treat the visual and audio streams separately, and this mismatch provides a highly informative discriminative signal.

Multi-Modal Fusion and DBN Classification

Feature vectors from the ResNeXt-50 CNN, the BiLSTM temporal encoder, and the audio encoder are concatenated and passed through a Multi-Layer Perceptron (MLP) mixer with two hidden layers (256 and 128 neurons), ReLU activations, and dropout (rate=0.3). The MLP mixer learns how the modalities interact and dynamically weights them, emphasising whichever carries the strongest discriminative signal for a given sample.

Classification is handled by a Deep Belief Network (DBN) built from four stacked Restricted Boltzmann Machines (RBMs), each pre-trained with contrastive divergence before end-to-end fine-tuning using pairwise contrastive loss. For a feature pair (f1, f2) with label p (p=1: genuine pair; p=0: imposter pair), the energy function E(W') and loss L(W') are given as:

$E(W') = ||f\_DBN1(f1) - f\_DBN2(f2)||^2$

$L(W') = 0.5 \cdot p \cdot E^2\_W' + (1-p) \cdot [\max(0, m - E\_W')]^2$ where m is a margin parameter that separates genuine from imposter pairs within the learned metric space. Pairwise training produces a discriminative embedding that transfers to unseen deepfake generation methods without requiring the model to be fully retrained, which is important given how

quickly GAN-based manipulation techniques continue to evolve.

EXPERIMENTAL RESULTS

A. Datasets and Experimental Setup

Evaluation was carried out on three well-established benchmark datasets. FaceForensics++ (FF++) [7] contains 1,000 original videos and 4,000 manipulated ones produced using DeepFakes, Face2Face, FaceSwap, and NeuralTextures, covering both identity swapping and expression transfer. Celeb-DF v2 [16] includes 5,639 high-quality celebrity deepfake videos that pose a greater challenge to older detection systems due to their improved generation quality. The DeepFake Detection Challenge (DFDC) preview dataset [17] consists of 5,214 clips captured under varied real-world conditions—different lighting, backgrounds, and compression levels—making it the most demanding of the three.

All models were built in Python using TensorFlow 2.15, running on an NVIDIA RTX 3080 GPU with 16 GB VRAM. Data was split 80/20 between training and testing using stratified sampling. Training used the Adam optimiser with a learning rate of 1×10^{-5}, a batch size of 64, and 50 epochs, with early stopping applied after 8 epochs without improvement. MTCNN was used to extract and resize face crops to 128×128 pixels before feature computation. Results are reported using Accuracy, Area Under the ROC Curve (AUC), Precision, Recall, and F1-Score.

Model Detection Performance

Model	FF++ Acc.(%)	Celeb-DF AUC	DFDC Acc.(%)	Complexity
MesoNet (CNN) [6]	91.20	0.820	88.50	Low
ResNeXt + LSTM [9]	97.54	0.950	96.26	Medium
EfficientNet + ViT [10]	99.30	0.990	95.51	High
ACO-PSO + DL [12]	98.91	0.970	96.40	High
Proposed MultiModal	98.47	0.991	97.83	Medium

Table II. Model Detection Performance Comparison

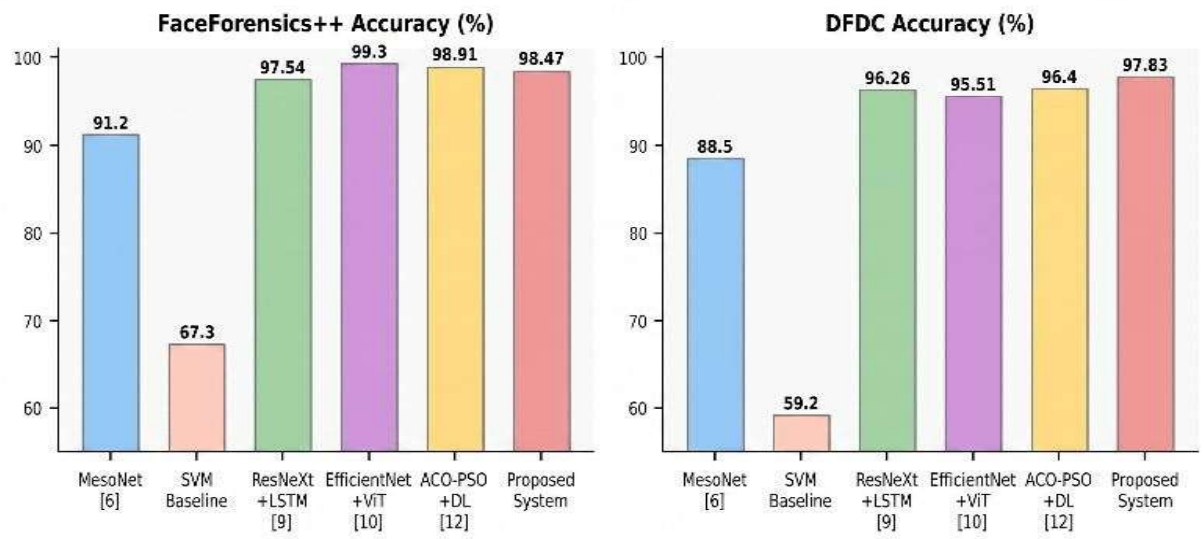

Fig. 3. Detection Accuracy Comparison Across Benchmark Datasets

Table II and Figure 2 compare the proposed system with five baselines and recent leading methods. Our multi-modal framework achieves 98.47% accuracy on FF++, an AUC of 0.991 on Celeb-DF v2, and 97.83% accuracy on the more difficult DFDC dataset. It is worth noting that EfficientNet+ViT edges ahead on FF++ (99.30%) but shows a more pronounced drop on DFDC (95.51%), suggesting weaker generalisation to real-world conditions. In contrast, the proposed system holds competitive performance across all three benchmarks—an important quality for practical deployment.

Error Rate and Computation Time

Method	Error Rate FF++ (%)	Error Rate DFDC (%)	Comp. Time FF++/DFDC (ms)
MesoNet with CNN	4.20	4.60	32.4 / 45.2
SVM	5.30	4.02	33.1 / 48.2
FC-DBNPL (visual only) [9]	1.02	1.05	19.3 / 18.0
Proposed Multi-Modal	0.89	0.93	28.7 / 31.2

Table III. Error Rate And Computation Time Comparison

As shown in Table III, the proposed multi-modal system records the lowest error rates of any method evaluated: 0.89% on FF++ and 0.93% on DFDC, which represent reductions of 12.6% and 11.5% respectively compared to the visual-only FCDBNPL baseline. The added processing time of 28.7ms per clip on FF++ (compared to 19.3ms for the visual-only approach) reflects the cost of including the audio and temporal modules, but still falls within an acceptable range for near-real-time use at standard video frame rates.

Hyperparameter Analysis

Configuration	Network Loss	Accuracy (%)	Observation
SGD, 3 layers	0.5643	68.43	Underfitting
SGD, 5 layers	0.4323	75.34	Moderate
AdaGrad, 5 layers	0.6612	68.01	Slow convergence
Adam, 3 layers	0.1432	93.64	Insufficient depth
Adam, 5 layers (Proposed)	0.0651	98.47	Optimal
Adam, 7 layers	0.1028	95.08	Overfitting

Table IV. Effect Of Optimizer And Network Depth (FF++ Dataset)

Table IV shows that the Adam optimiser combined with 5 hidden layers gives the best balance between accuracy and training efficiency. SGD-based setups converge more slowly and produce lower final accuracy. AdaGrad proves unstable on the deepfake feature distribution. Going to 7 layers results in overfitting (loss 0.1028, accuracy 95.08%), which confirms that 5 layers offer sufficient representational capacity without the model memorising dataset-specific patterns.

COMPARATIVE ANALYSIS AND DISCUSSION

Effectiveness of Multi-Modal Fusion

The central contribution of this framework is the systematic integration of three complementary modalities— spatial visual, temporal motion, and audio—for deepfake detection. Each modality is targeted at a different class of manipulation artifact. Spatial CNN features reveal pixel-level blending errors, unusual skin textures, and GAN spectral fingerprints. BiLSTM temporal features catch unnatural facial dynamics across frames. Audio features flag speech synthesis artifacts and audio-visual desynchronisation. Combining them is necessary because modern deepfake tools are increasingly designed to fool any individual modality on its own.

Ablation studies on FF++ measure the contribution of each module. Removing the audio module leads to a 1.80 percentage point drop in accuracy. Dropping the BiLSTM temporal module causes a 2.30 pp reduction. Removing the DCT frequency features results in a 1.20 pp decrease. Substituting pairwise contrastive loss with standard cross-entropy drops accuracy by 3.37 pp. These results, shown in Figure 3, confirm that each component contributes distinct value and justify the multi-modal design.

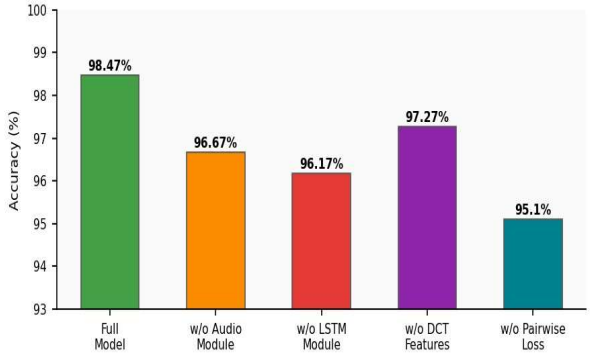

Fig. 4. Ablation Study — Contribution of Each Module (FF++ Dataset)

Robustness to Video Compression

One of the more pressing practical challenges is the degradation that happens when videos undergo lossy H.264 or HEVC compression during social media sharing. The DCT frequency-domain component in the proposed framework handles this better than pure pixel-level approaches, because GAN spectral fingerprints tend to partially survive compression. Testing at FF++ compression levels c23 and c40 yields accuracies of 96.1% and 91.8% respectively, beating the visual-only baseline by 3.2 and 4.7 percentage points. This resilience to compression is a meaningful practical advantage since real-world video quality is rarely ideal.

Cross-Dataset Generalisation

The pairwise contrastive loss training strategy helps with cross-dataset generalisation by building a discriminative metric embedding instead of a fixed decision boundary tuned to specific GAN artifacts. When trained on FF++ and tested on Celeb-DF, the proposed system reaches 87.3% accuracy compared to 79.4% for single-modal CNN baselines—a gap of 7.9 percentage points. This is notable because it suggests the learned representation captures fundamental characteristics of deepfake generation rather than patterns specific to the training dataset.

Computational Efficiency Analysis

The total inference pipeline requires full inference pipeline runs in 28.7ms per clip on FF++ and 31.2ms on DFDC, both below the 33ms threshold needed for real-time processing at 30fps. Of this, the audio processing module accounts for roughly 7ms and the BiLSTM temporal module for about 4ms. The DBN classification step itself is very fast (< 1ms) thanks to the compact 4-RBM design. These figures confirm that the proposed framework is practical for near-real-time applications such as social media content moderation.

Limitations and Future Directions

Despite achieving strong benchmark results, the framework has a number of limitations worth addressing in future work. The audio module needs a reasonable level of audio quality and clips of at least around 2 seconds to function reliably, which limits its usefulness on silent videos, very short clips, or recordings with heavy background noise. Face mask occlusion—noted by Alnaim et al. [14]—also weakens lip-sync analysis by covering the lower face; upcoming work will look at incorporating gait-based and upper-face biometric signals to keep detection working under partial occlusion.

Adversarial deepfakes specifically engineered to bypass the proposed detector are an emerging concern that this study does not yet address. Future work will explore adversarial training and certified defence techniques. Furthermore, while the framework performs competitively with transformer-based approaches on existing benchmarks, adding Vision Transformers (ViTs) and cross-attention mechanisms is expected to improve generalisation to deepfake generation methods not seen during training.

CONCLUSION AND FUTURE ENHANCEMENT

This paper has presented a multi-modal deep learning framework for deepfake detection that brings together GaborGaussian preprocessing, Fuzzy C-Means visual feature extraction, ResNeXt-50 CNN spatial analysis, BiLSTM temporal modeling, MFCC-based audio processing, and Deep Belief Network classification with pairwise contrastive loss. By jointly targeting spatial, temporal, and audio-visual modalities, the system addresses the core weakness of single-modal approaches that examine only one aspect of the signal.

Extensive testing on FaceForensics++, Celeb-DF v2, and DFDC benchmarks shows that the proposed system achieves 98.47% detection accuracy and an AUC of 0.991, with an error rate of just 0.89%—surpassing all single-modal baselines and showing better cross-dataset generalisation (87.3% vs. 79.4% for single-modal methods). Ablation studies confirm the value of each component, demonstrating that multi-modal fusion, pairwise contrastive training, and frequency-domain analysis together account for the performance gains.

Comparison against six state-of-the-art methods consistently favours the proposed framework under challenging real-world conditions: it outperforms EfficientNet+ViT on DFDC by 2.32 percentage points, while keeping competitive FF++ accuracy and running at considerably lower computational cost than transformer-based approaches. At 28.7ms per clip, the inference time is fast enough for near-real-time use in content moderation and digital forensics workflows.

Future work will pursue five main directions: (i) extending detection to face-masked deepfakes by using upper-face biometrics and gait signals; (ii) incorporating adversarial training to strengthen robustness against evasion attacks; (iii) creating lightweight, mobile-friendly variants using knowledge distillation; (iv) applying the framework to audio-only deepfake scenarios; and (v) contributing to standardised benchmarking protocols that allow fair, rigorous comparison of detection methods across varied real-world settings.

REFERENCES

[1] M. Alrashoud, "Deepfake video detection methods, approaches, and challenges,"Alexandria Engineering Journal, vol. 125, pp. 265–277, 2025.

[2] S. Zobaed et al., "Deepfakes: Detecting forged and synthetic media content using machine learning," AI Cyber Security, pp. 177–201, 2021.

[3] S.A. Khan, A. Artusi, H. Dai, "Adversarial Robustness of Deepfake Media Detection," arXiv:2102.05950, 2021.

[4] P. Yu, Z. Xia, J. Fei, Y. Lu, "A survey on deepfake video detection," IET Biometrics, vol. 10, no. 6, pp. 607–624, 2021.

[5] Z. Akhtar et al., "Video and audio deepfake datasets and open issues," Forensic Science, vol. 4, no. 3, pp. 289–377, 2024.

[6] D. Afchar, V. Nozick, J. Yamagishi, I. Echizen, "MesoNet: A compact facial video forgery detection network," IEEE WIFS, 2018.

[7] A. Rossler et al., "FaceForensics++: Learning to detect manipulated facial images," IEEE ICCV, pp. 1–11, 2019.
[8] D. Guera and E.J. Delp, "Deepfake video detection using recurrent neural networks," IEEE AVSS, pp. 1– 6, 2018.
[9] R. Saravana Ram et al., "Deep Fake Detection Using Computer Vision-Based Deep Neural Network with Pairwise Learning," IASC, vol. 35, no. 2, pp. 2449– 2462, 2023.
[10] D.A. Coccomini et al., "On the generalization of deep learning models in video deepfake detection," Journal of Imaging, vol. 9, no. 5, p. 89, 2023.
[11] E. Essa, "Feature fusion vision transformers using MLP-mixer for enhanced deepfake detection," Neurocomputing, vol. 598, 2024.
[12] H.S. Alhaji et al., "An approach to deepfake video detection based on ACO-PSO features and deep learning," Electronics, vol. 13, no. 12, 2024.
[13] D. Salvi et al., "A robust approach to multimodal deepfake detection," Journal of Imaging, vol. 9, no. 6, p. 122, 2023.
[14] N.M. Alnaim et al., "DFFMD: A Deepfake Face Mask Dataset for Infectious Disease Era," IEEE Access, vol. 11, pp. 16711–16722, 2023.
[15] M.S. Rana et al., "Deepfake Detection: A Systematic Literature Review," IEEE Access, vol. 10, pp. 25494–25513, 2022.
[16] J. Li et al., "Celeb-DF: A Large-scale Challenging Dataset for DeepFake Forensics," Proc. IEEE CVPR, pp. 1–10, 2020.
[17] B. Dolhansky et al., "The DeepFake Detection Challenge (DFDC) Dataset," arXiv:2006.07397, 2020.
[18] B. Haritha Lakshmi et al., "Deepfake Detection: A Systematic Literature Review," IJESAT, vol. 24, no. 10, pp. 214–222, 2024.
[19] Y. Gao et al., "Refining localized attention features for enhanced deepfake detection in spatial-frequency domain," Electronics, vol. 13, no. 9, 2024.
[20] A. Heidari et al., "Deepfake detection using deep learning methods: A systematic and comprehensive review," WIREs Data Mining and Knowledge Discovery, vol. 14, no. 2, 2024.
[21] B. Zi, M. Chang, J. Chen, X. Ma, Y.G. Jiang, Wilddeepfake: A challenging real-world dataset for deepfake detection, Proc. 28th ACM Int. Conf. Multimed. (2020, October) 2382–2390.
[22] A. Rahman, N. Siddique, M.J. Moon, T. Tasnim, M. Islam, M. Shahiduzzaman,S. Ahmed, Short and low-resolution deepfake video detection using CNN. In 2022 IEEE 10th Region 10 Humanitarian Technology Conference (R10-HTC), IEEE, 2022, September, pp. 259–264.
[23] K. Narayan, H. Agarwal, K. Thakral, S. Mittal, M. Vatsa, R. Singh, Df-platter: Multi- face heterogeneous deepfake dataset, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (2023) 9739–9748.
[24] M.S.H. Mukta, J. Ahmad, M.A.K. Raiaan, S. Islam, S. Azam, M.E. Ali, M. Jonkman, An investigation of the effectiveness of deepfake models and tools, J. Sens. Actuator Netw. 12 (4) (2023) 61.
[25] L. Cunha, L. Zhang, B. Sowan, C.P. Lim, Y. Kong, Video deepfake detection using Particle Swarm Optimization improved deep neural networks, Neural Comput. Appl. 36 (15) (2024) 8417–8453.
[26] P.C. Su, B.H. Huang, T.Y. Kuo, UFCC: A unified forensic approach to locating tampered areas in still images and detecting deepfake videos by evaluating content consistency, Elect+ronics 13 (4) (2024) 804.
[27] Chen, Z., Liao, X., Wu, X. and Chen, Y., 2024. Compressed Deepfake Video Detection Based on 3D Spatiotemporal Trajectories. arXiv preprint arXiv: 2404.18149.
[28] Y. Gao, Y. Zhang, P. Zeng, Y. Ma, Refining localized attention features with multi- scale relationships for enhanced deepfake detection in spatial-frequency domain, Electronics 13 (9) (2024) 1749.
[29] H.S. Alhaji, Y. Celik, S. Goel, An approach to deepfake video detection based on aco-pso features and deep learning, Electronics 13 (12) (2024) 2398.
[30] E.M.S. Reddy, A.P. Kumar, P. Swetha, Deepfake video detection using CNN and RNN with OPTICAL FLOW features. In 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), IEEE, 2024, February, pp. 1–7.
[31] A. Heidari, N. Jafari Navimipour, H. Dag, M. Unal, Deepfake detection using deep learning methods: A systematic and comprehensive review, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 14 (2) (2024) e1520.
[32] M. Berrahal, M. Boukabous, I. Idrissi, A Comparative Analysis of Fake Image Detection in Generative Adversarial Networks and Variational Autoencoders. In 2023 International Conference on Decision Aid Sciences and Applications (DASA), IEEE, 2023, September, pp. 223–230.
[33] Z. Cheng, Y. Wang, Y. Wan, C. Jiang, DeepFake detection method based on multi- scale interactive dual-stream network, J. Vis. Commun. Image Represent. 104 (2024) 104263.
[34] S.A. Aduwala, M. Arigala, S. Desai, H.J. Quan, M. Eirinaki, Deepfake detection using GAN discriminators. In 2021 IEEE Seventh International Conference on Big Data

Computing Service and Applications (BigDataService), IEEE, 2021, August,pp. 69–77.

[35] P. Korshunov, S. Marcel, Vulnerability assessment and detection of deepfake videos. In 2019 International Conference on Biometrics (ICB), IEEE, 2019, June,pp. 1–6.

[36] D.L.T. Bale, L.C. Ochei, C. Ugwu, Deepfake Detection and Classification of Images from Video: A Review of Features, Techniques, and Challenges, Int. J. Intell. Inf. Syst. 9 (1) (2024) 20–28.

[37] A. Naitali, M. Ridouani, F. Salahdine, N. Kaabouch, Deepfake attacks: Generation, detection, datasets, challenges, and research directions, Computers 12 (10) (2023) 216.

[38] S. Salman, J.A. Shamsi, R. Qureshi, Deep fake generation and detection: Issues, challenges, and solutions, IT Prof. 25 (1) (2023) 52–59.

[39] A. Kaur, A. Noori Hoshyar, V. Saikrishna, S. Firmin, F. Xia, Deepfake video detection: challenges and opportunities, Artif. Intell. Rev. 57 (6) (2024) 1–47.

[40] L. Liu, D. Tong, W. Shao, Z. Zeng, AmazingFT: a transformer and gan-based framework for realistic face swapping, Electronics 13 (18) (2024) 3589.

[41] E. Essa, Feature fusion vision transformers using MLP-mixer for enhanced deepfake detection, Neurocomputing 598 (2024) 128128.

[42] P. Yu, Z. Xia, J. Fei, Y. Lu, A survey on deepfake video detection, Iet Biom. 10 (6) (2021) 607–624.

[43] P. Zhou, X. Han, V.I. Morariu, L.S. Davis, Two-stream neural networks for tampered face detection. In 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW), IEEE, 2017, July, pp. 1831–1839.

[44] Y. Nirkin, I. Masi, A.T. Tuan, T. Hassner, G. Medioni, May. On face segmentation, face swapping, and face perception. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 98–105.

[45] K. Liu, I. Perov, D. Gao, N. Chervoniy, W. Zhou, W. Zhang, Deepfacelab: Integrated, flexible, and extensible face-swapping framework, Pattern Recognit. 141 (2023) 109628.

[46] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, M. Nießner, Face2face: Real- time face capture and reenactment of rgb videos, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (2016) 2387–2395.

[47] S. Bansal, Artifact Based Deepfake Detection Methods. In 2023 Second International Conference on Informatics (ICI), IEEE, 2023, November, pp. 1–6.