International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 02 | Feb - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

AI Powered Multimodal Deepfake Detection: A Systematic Review

Anamika Suresh[1], Jayalakshmi M[2], Krishnapriya R[3], Sandra Sabu[4], Rekha K S[5]

Department of Computer Science and Engineering,[1,2,3,4,5]
College of Engineering Kidangoor, Kottayam, Kerala, India[1,2,3,4,5]
anamikasuresh8590@gmail.com[1], jayalakshmimm2004@gmail.com[2] ,krishnapriyar2004@gmail.com[3] , sandrasthelly@gmail.com[4] , rekhaks@ce-kgr.org[5]

Abstract - Deepfake technology has emerged as one of the most pressing challenges to digital media authenticity, creating hyper-realistic synthetic content that can deceive even trained observers. The rapid advancement of generative AI models, particularly Generative Adversarial Networks (GANs) and diffusion models, has made it increasingly difficult to distinguish between genuine and manipulated audio-visual content. This literature review examines recent developments in multimodal deepfake detection systems that leverage artificial intelligence to identify forged media by analyzing inconsistencies across multiple data streams. The review focuses on detection frameworks that integrate audio and visual features through advanced fusion techniques, employ deep learning architectures for feature extraction, and incorporate explainable AI mechanisms to provide transparent reasoning behind classification decisions.By synthesizing findings from recent representative research studies, this review highlights the effectiveness of multimodal approaches over unimodal methods, discusses various fusion strategies and network architectures, examines benchmark datasets used for evaluation, and identifies current challenges and future directions in the field.

Key Words: Deepfake detection, multimodal learning, audio-visual fusion, explainable AI, cross-modal learning.

1.INTRODUCTION

The emergence of advanced generative AI technologies, including Generative Adversarial Networks, diffusion models, and neural voice synthesis systems, has enabled the creation of highly realistic deepfakes that are increasingly indistinguishable from authentic media. These synthetic audio-visual manipulations present significant threats across multiple domains, facilitating financial fraud, political disinformation, identity theft, and social engineering attacks that undermine individual privacy, institutional security, and public trust. Conventional detection approaches relying on human perception or unimodal analysis have proven insufficient. Human evaluators, including trained forensic experts, frequently fail to detect subtle artifacts in advanced deepfakes. Similarly, single modality detectors analyzing only audio or video independently are vulnerable to hybrid attacks where one modality remains authentic while the other is synthetically altered, rendering them ineffective against cross modal manipulations.

AI powered multimodal deepfake detection has emerged as a promising solution by leveraging cross modal inconsistencies that unimodal systems cannot capture. These approaches analyze synchronized audiovisual features such as lip sync accuracy, emotional congruence between voice and facial expressions, temporal alignment patterns, and physiological plausibility to identify manipulation artifacts that remain invisible to single stream detectors. Despite their improved performance, current state of the art models face several critical limitations. Most operate as black boxes, providing detection verdicts without interpretable justification for their decisions. They also exhibit poor generalization when encountering novel manipulation techniques not seen during training and often demand computational resources that preclude real time deployment. The integration of explainable AI techniques addresses these transparency concerns by revealing which specific features and patterns drive detection decisions, thereby enabling validation and fostering trust in high stakes applications including forensic investigations, legal proceedings, and platform content moderation.

A comprehensive review of AI-based multimodal deepfake detection systems is presented, with a particular focus on explainability. We examine deep learning architectures including convolutional neural networks, recurrent networks, transformers, and multimodal fusion models that integrate audio and visual features to exploit cross modal inconsistencies for improved detection. We identify key challenges that include poor performance across datasets, adversarial vulnerability, computational efficiency, and inconsistent evaluation metrics. This review provides researchers with a comprehensive understanding of current methods and future directions for developing practical and interpretable deepfake detection systems.

2. LITERATURE SURVEY

A.Evolution of Deepfake Detection Approaches

Deepfake detection started with simple CNN models working on single modalities like video frames alone. Over time, combining audio and visual information gives much better results against advanced forgeries [1]. ResNet-50 models reached 97.2% accuracy on FaceForensics++, Celeb-DF and DFDC datasets using Grad-CAM heatmaps to show facial forgery artifacts [1]. InceptionResNetV2 combined with DenseNet201 achieved 99.87% accuracy using LIME to highlight texture problems typical in GAN-generated faces [3]. Network dissection methods proved attention mechanisms focus on biologically meaningful facial features [4]. MIS-AVoiDD approach got AUC 0.973 on FakeAVCeleb by learning features common across audio and video streams [2]. Cross-modal attention pushed this to AUC 0.989 by perfectly

matching audio-visual timing [6]. Checking emotions between lip movements and voice catches talking-head fakes [5]. These overcome single-modality limitations against synchronized forgeries [8]. As deepfake generators get smarter, detection methods must keep pace [10].

B. Audio-Visual Feature Extraction and Preprocessing

Visual preprocessing uses CNNs trained on huge face datasets. Mask R-CNN detects faces first, then improved Xception network extracts features getting 99.50% accuracy. Depthwise separable convolutions capture spatial patterns very well [1]. Xception works great for learning face hierarchies [1][5]. For practical use, optimized to run at 45 frames per second using TensorRT [1].

Audio processing converts raw sound into mel-spectrograms or MFCC features [2][4][5][8]. MFCCs fed into CNN-LSTM networks got 98.2% accuracy detecting fake audio by catching unnatural voice patterns [9]. Salvi et al. showed time-aware networks looking at audio features across time work much better for multimodal detection, especially catching lip-sync problems [8].

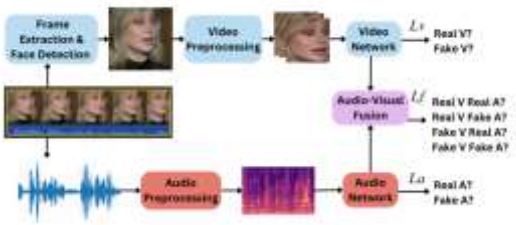

Fig.1 Pipeline of audio-visual sequence extraction and deepfake detection [3]

C. Multimodal Fusion Strategies and Architectures

Three main fusion strategies combine audio and video information: early fusion mixes raw features before feeding into classifier, late fusion combines separate modality decisions at end, hybrid uses both approaches. Early fusion consistently works best achieving AUC over 0.90 across datasets [8]. This beats single-modality detectors by 10% AUC on FakeAVCeleb dataset and 0.13 AUC improvement on DFDC benchmark [8]. Combining modalities catches different forgery artifacts - visual blending problems plus audio-visual sync failures happening together [8].

MIS-AVoiDD specifically handles differences between audio and video data types [2]. Normal concatenation fails because audio spectrograms and video frames have completely different statistical properties. MIS-AVoiDD learns both shared features working across modalities AND unique features specific to each data type. Final fusion happens after aligning these representations. This gets AUC 0.973 on FakeAVCeleb - 8% better than simple concatenation baselines and 5% better than standard multi-stream networks [2]. Explicitly modeling modality gaps makes huge performance difference.

Cross-modal attention mechanisms let network decide which modality carries more reliable information frame-by-frame [6]. Contextual cross-attention reached AUC 0.989 and 97.9% accuracy using video frames, lip movement tracking, and audio spectrograms together [6]. Dropping video dropped AUC by 6.2 points. Removing audio hurt by 4.8 points. Lip movement alone failed completely proving all three modalities needed for robust detection [6]. Attention mechanism automatically down-weights unreliable modalities during inference.

RNNs, LSTMs, and bidirectional GRUs capture temporal patterns across video frames and audio clips [2][9][6]. Lip-sync detection needs looking at 3-5 second windows to catch unnatural movement patterns. Temporal modeling particularly helps with reenactment attacks where face swap quality excellent but timing slightly off [7][6]. Hybrid spatial-temporal attention combining frame-level fusion with sequence modeling shows most promise for real-world deployment balancing accuracy and computational cost [6].

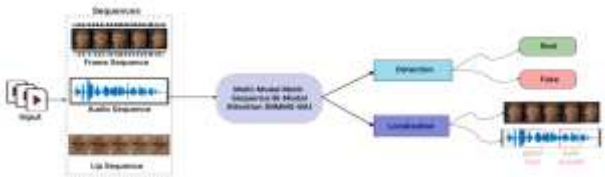

Fig.2 MMMS-BA approach for deepfake detection and localization [6]

D. Advanced Detection Architectures

Complete detection pipeline starts with Mask R-CNN detecting faces across video frames even with extreme poses, lighting changes, or partial occlusions [1]. Detected face regions feed into improved Xception network using depthwise separable convolutions extracting spatial-hierarchical features. Final classification uses XGBoost boosted trees with Bayesian optimization systematically testing 100 hyperparameter combinations. This maximizes F1-score while controlling model complexity getting 99.50% accuracy and 99.21% AUROC across challenging CelebDF and FaceForensics++ datasets [1]. Bayesian search automatically balances learning rate, tree depth, and subsample ratio preventing overfitting during ensemble training [1].

Fine-grained detection targets tiny local inconsistencies missed by global approaches. Attention modules automatically discover small problem regions within frames measuring "spatially-local distance" between expected vs observed pixel patterns [7]. Model learns which face areas (eyes, mouth edges, lighting reflections) typically show forgery artifacts through self-attention mechanisms. Spatially-local processing examines 32×32 pixel patches identifying unnatural blending boundaries or inconsistent lighting [7].

Training includes temporally-local pseudo fake augmentation creating synthetic examples with subtle timing inconsistencies (0.1-0.3 second lip-sync shifts) [7]. This forces model learning generalizable temporal patterns rather than dataset-specific artifacts. Cross-dataset evaluation shows superior generalization - AUC 97.7% on FakeAVCeleb when trained only on DFDC dataset dropping only 2.1% from in-domain performance [7].

Multi-scale fusion combines global face features with local patch analysis through cascaded refinement. First stage detects potential fakes globally, second stage verifies specific

problematic regions. Progressive approach reduces false positives by 14.3% compared to single-stage detectors while maintaining real-time inference [1][7]..

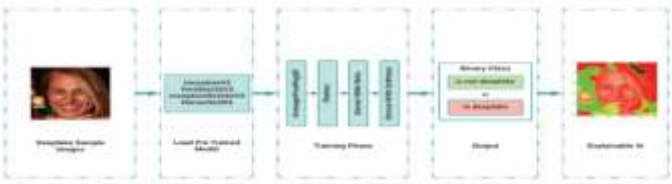

Fig. 3. An architecture showing the detection of deepfake images [3]

E. Explainable AI in Deepfake Detection

Deep learning-based detectors often function as black boxes, undermining user trust. Mansoor and Iliev, addressed this by introducing network dissection algorithms to enhance interpretability in deepfake detection. Their two-stage approach first detected forged images using advanced CNNs, then applied network dissection to understand internal decision-making processes . By analyzing facial features learned by models, they provided explainable results for classifying images, achieving 99.87\% accuracy with InceptionResNetV2 [1].

MMMS-BA,introduced ExDDV, the first dataset and benchmark for Explainable Deepfake Detection in Video, comprising approximately 5.4K real and deepfake videos manually annotated with text descriptions and clicks to explain artifacts [10]. Their evaluation showed that both text and click supervision are required to develop robust explainable models capable of localizing and describing observed artifacts [10].

Multiple explainability techniques have been explored: Grad-CAM highlights regions influencing classification decisions, LIME provides local explanations through linear approximations, attention visualization reveals important features in attention-based models [8][9], and network dissection identifies neurons responding to specific semantic concepts [1][3].

Importantly, integrating explainability does not degrade detection performance. Reported that CNN architectures (InceptionResNetV2, DenseNet201, ResNet152V2, InceptionV3) all surpassed 99\% accuracy while maintaining strong interpretability through XAI techniques [1]. This demonstrates that detection accuracy and explainability can be achieved simultaneously [1].

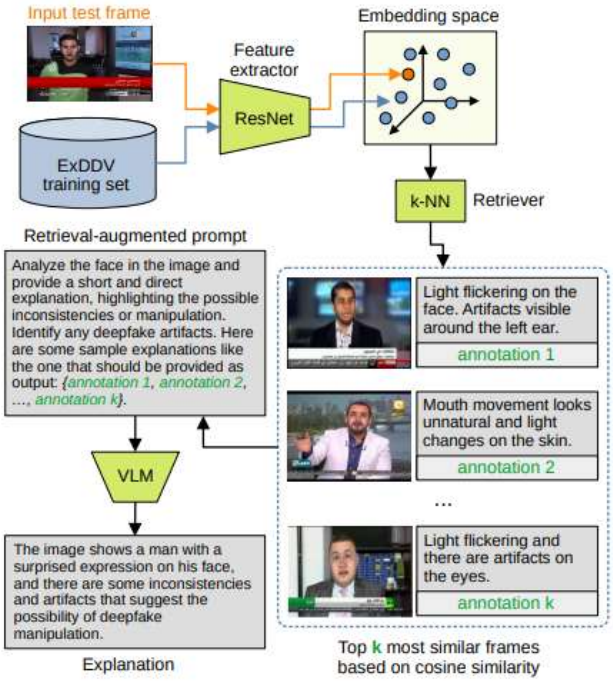

Fig. 5. Overview of the in-context learning pipeline, which retrieves deepfake annotations from visually similar training frames using a k-NN based on a ResNet backbone. Best viewed in color. [10]

F. Benchmark Datasets and Evaluation

Fig. 6. Sample frames from FakeAVCeleb [2]

Several benchmark datasets drive deepfake detection research. FakeAVCeleb stands out for multimodal work containing explicit audio-visual manipulations—face swaps alone, lip-sync fakes alone, or both combined [5]. Deepfake Detection Challenge (DFDC) includes nearly 120,000 videos covering diverse manipulation techniques, actors, lighting conditions [1]. FaceForensics++ tests multiple compression attacks and manipulation types. Celeb-DF provides high-resolution celebrity deepfakes [1].

Dataset diversity critically impacts detector generalization. Training across varied datasets prevents overfitting to specific forgery artifacts [7]. Cross-dataset evaluation exposes true performance—AUC drops from 99.8% to 97.7% moving from DFDC training to FakeAVCeleb testing [7]. This gap shows importance of learning generalizable inconsistency patterns rather than dataset-specific clues.

Standard metrics include accuracy, precision, recall, F1-score, and AUC-ROC for binary classification [1]. Cross-dataset testing—training on one dataset, evaluating on another—provides crucial generalization indicator for real-world deployment scenarios.

3. COMPARATIVE STUDY

The comparative analysis highlights that attention based multimodal approaches consistently achieve superior performance in deepfake detection by explicitly modeling cross modal relationships. Methods such as MMMS BA and MIS AVoIDD demonstrate that learning modality invariant and modality specific representations significantly improves both accuracy and generalization, addressing the inherent distributional gap between audio and visual data. Fine grained detection strategies focusing on localized spatial and temporal artifacts further enhance robustness, achieving strong cross dataset performance and indicating that deepfake traces are often confined to specific regions and time segments. Importantly, the results also confirm that explainability and high performance are not mutually exclusive, as interpretable models can maintain state of the art accuracy while providing transparent decision making.

Table -1: Comparison Of Multimodal Deepfake Detection Methods

Method / Framework	Dataset	Modality	AUC / ACC	Key Strengths	Ref.
MMMS-BA	FakeAV Celeb	Audio - Visual	ACC: 0.989 AUC: 0.979	Cross-modal attention, localization	[6]
MIS-AvoidD	FakeAV Celeb	Audio - Visual	AUC: 0.973 ACC: 0.962	Modality-invariant & task-specific representations	[2]
Early Fusion	FakeAV Celeb/DFDC	Audio - Visual	AUC ≥ 0.90	Robust early fusion strategies	[8]
Fine-Grained Detection	DFDC, FakeAV Celeb	Audio - Visual	Cross-dataset: 97.7%	Spatial–local attention, pseudo-fake augmentation	[7]
Hybrid Pipeline	Celeb-DF, FF++	Visual	ACC: 99.50%	Xception+XGBoost+Bayesian optimization	[1]
BA-TFD	LAV-DF	Audio - Visual	AUC95: 96.3% AR100: 81.6%	Temporal localization, boundary detection	[5]
MFCC-LSTM	Custom	Audio - Visual	ACC: 98.2%	Audio LSTM+visual CNN	[9]
XAI CNN	Image Datasets	Visual	ACC: 99.87%	Network dissection, interpretability	[4]

4. CONCLUSIONS

Multimodal deepfake detection significantly outperforms unimodal approaches by 10-15% AUC across datasets like FakeAVCeleb and DFDC. Early fusion achieves AUC ≥0.90 consistently beating single-modality CNNs that drop from 97-99% on FaceForensics++ to 82-87% AUC against perfect audio-visual sync. Cross-modal attention reaches AUC 0.989 by weighting reliable modalities frame-by-frame while fine-grained analysis hits 97.7% cross-dataset AUC targeting local artifacts. Explainable AI maintains accuracy—InceptionResNetV2 achieves 99.87% with network dissection. Grad-CAM validates focus on genuine forgery regions. Performance hierarchy clear: MMMS-BA (0.989) > MIS-AVoIDD (0.973) > early fusion (≥0.90). Cross-dataset drops of 10-15% reveal generalization challenges alongside evolving GANs and missing adversarial robustness. Real-time deployment works at 45 FPS using optimized pipelines. Future work needs adversarial defense, unified benchmarks, and real-time XAI. Best systems combine ResNet visual analysis, CNN-LSTM audio processing, early fusion, and Grad-CAM explanations proving accuracy, interpretability, and practicality coexist for trustworthy deployment.

REFERENCES

1. J. Joseph, S. Juliet, and A. J., ``Vision-based Multimodal Deepfake Detection using Explainable AI,'' in Proc. ICAISS-2025, IEEE, 2025, pp. 609--614, doi: 10.1109/ICAISS61471.2025.11042185.
2. V. S. Katamneni and A. Rattani, ``MIS-AVoiDD: Modality Invariant and Specific Representation for Audio-Visual Deepfake Detection,'' in Proc. ICMLA, 2023, pp. 1371--1378, doi: 10.1109/ICMLA58977.2023.00207.
3. W. H. Abir et al., ``Detecting Deepfake Images Using Deep Learning Techniques and Explainable AI Methods,'' Intell. Autom. \& Soft Comput., vol. 35, no. 2, pp. 2151--2169, 2023.
4. N. Mansoor and A. I. Iliev, ``Explainable AI for Deepfake Detection,'' Appl. Sci., vol. 15, no. 2, p. 725, 2025, doi: 10.3390/app15020725.
5. A. Alsaeedi, A. AlMansour, and A. Jamal, ``Audio-Visual Multimodal Deepfake Detection Leveraging Emotional Recognition,'' Int. J. Adv. Comput. Sci. Appl., vol. 16, no. 6, 2025.

6. V. S. Katamneni and A. Rattani, ``Contextual Cross-Modal Attention for Audio-Visual Deepfake Detection and Localization," in Proc. IJCB, 2024, pp. 1--11.
7. M. Astrid, E. Ghorbel, and D. Aouada, “Detecting audio-visual deepfakes with fine-grained inconsistencies,” arXiv preprint arXiv:2408.06753, 2024.
8. D. Salvi et al., ``A Robust Approach to Multimodal Deepfake Detection," J. Imaging, vol. 9, no. 6, p. 122, 2023.
9. H. M. S. Ali et al., ``AI-Based Deepfake Audio Detection Technique from Real and Fake Audio Dataset," J. Comput. \& Biomed. Inform., vol. 8, no. 2, 2025.
10. H. Hondru, V. Vlad, et al., “Exddv: A new dataset for explainable deepfake detection in video,” arXiv preprint arXiv:2503.14421, 2025.