RESEARCH ARTICLE

Intelligent Deepfake Detector Using Audio-Visual Clues

Barnali Gupta Banik[1,*] **and Shaik Nidha Naziya**[2]

[1] Mahatma Gandhi Institute of Technology, Hyderabad, Telangana, India
[2] Malla Reddy Engineering College for Women, Hyderabad, Telangana, India

Abstract

Deepfake media is growing rapidly and causing significant harm. Bad actors now use AI to create fake videos that appear increasingly realistic. Traditional detection tools often fail because they analyze audio or visual signals in isolation. This paper introduces an intelligent Deepfake Detection system that addresses this limitation through a novel Multi-Modal Dispersion Framework. The system identifies subtle inconsistencies by tracking how lip movements align with speech patterns. By projecting these features into a shared latent space, the model quantifies the semantic divergence between modalities. A transformer module then captures cross-modal context to detect fine-grained manipulation artifacts. Evaluated on the DFDC and FakeAVCeleb datasets, the system achieves 94.3% accuracy, demonstrating strong potential for real-time deployment. This framework provides a reliable approach to media authentication and contributes to advancing AI safety.

Keywords: deepfake detection, multi-modal dispersion, audio-visual clues, cross-modal inconsistency, lip-sync analysis, AI forensics, transformer fusion.

Submitted: 21 September 2025
Accepted: 20 January 2026
Published: 01 March 2026

Vol. 2, **No.** 2, 2026.
10.62762/TMI.2025.601369

***Corresponding author:**
✉ Barnali Gupta Banik
barnali.guptabanik@ieee.org

1 Introduction

Deepfake technology is evolving rapidly through generative models like GANs and diffusion networks [1–3]. These tools create fabricated videos that threaten digital trust and public security [4, 5]. Early detection methods usually focused on just one modality, such as visual artifacts or audio spoofing [15–17]. These unimodal tools often fail against modern, sophisticated forgeries [6, 7]. Visual detectors look for blinking issues or facial warping [13, 15]. Audio detectors track frequency distortions or synthesis glitches [16, 17]. Newer research shows that combining audio and visual data is much more effective [1, 2, 8]. Deepfakes often struggle to sync lip movements with spoken words or speech prosody [4, 5, 11]. Modern fusion frameworks now learn from both facial cues and acoustic features [3, 10, 19]. Specific methods use transformer models and attention mechanisms to find these subtle gaps [1, 2, 11]. Some frameworks also project features into a shared space to measure divergence using distance metrics [3, 12, 18]. Recent tests on DFDC and FakeAVCeleb datasets prove these multi-modal models outperform older ones [2, 10, 19]. Experts agree that the field needs more scalable and interpretable solutions [6, 7, 20]. This paper introduces an intelligent Deepfake Detection System based on a Multi-Modal Dispersion Framework. The system combines lip-syncing and speech patterns to find inconsistencies. It aligns these features in a common space to quantify their divergence. The model achieves over 93% accuracy in real-time

Citation
Banik, B. G., & Naziya, S. N. (2026). Intelligent Deepfake Detector Using Audio-Visual Clues. *ICCK Transactions on Machine Intelligence*, 2(2), 100–105.

scenarios. This work provides a robust solution for forensics and AI safety [9, 14, 20].

2 Related Work

Deepfake detection is a vital research area due to the spread of synthetic media in social and forensic fields [1, 4, 6]. Researchers have tested many methods, ranging from manual feature checks to advanced deep learning. Early work focused on visual cues like eye blinking, facial warping, and head pose errors [13, 15]. Li et al. [15] identified warping artifacts common in early GAN-based deepfakes. Other studies used spatial attention and convolutional backbones to find tiny texture flaws [13, 19]. However, these visual methods often fail when videos are compressed or look very realistic. Audio detection focuses on voice synthesis artifacts and pitch patterns using CNNs or recurrent models [16, 17]. Chintha et al. [17] built a structure to find both audio spoofing and deepfakes, while Masood et al. [16] analyzed countermeasures for audio threats. These methods can still struggle with high-quality audio or transferred real voices.

Multi-modal detection solves these issues by combining audio and visual clues [1–3, 10]. It exploits errors like lip-sync delays or mismatched phonemes. For example, AV-Lip-Sync+ uses multi-scale attention to find timing gaps between video and sound [1]. DF-TransFusion uses cross-attention between lips and audio to increase sensitivity [2]. Frameworks like M2TR and AVoiD-DF use transformers to align both inputs into one space [3, 11]. These models perform better on datasets like DFDC and FakeAVCeleb [4, 5]. Multi-modal dispersion maps these features into a shared space to measure their semantic distance [3, 12, 18]. This helps quantify how far audio and video deviate in fake content. Rossler et al. [18] expanded this for source attribution, while AVFakeNet used late fusion to help alignment [19]. Recent surveys by Nguyen-Le et al. [6] and Liu et al. [7] track the move toward these joint methods. Sharma et al. [10] also confirmed that multi-modal systems are superior to single-mode tools. The IEEE Future Directions Committee highlights that these systems are now essential for real-time media forensics and safety [9].

3 Methodology

This section presents the architecture and methodology of the intelligent Deepfake Detection System. It exploits cross-modal inconsistencies using the Multi-Modal Dispersion Framework (MMDF). The system consists of five main components as shown in Figure 1.

3.1 Audio and Visual Feature Extraction

The framework uses advanced neural encoders to capture fine-grained patterns. The visual branch samples video frames at 25 fps and uses a 3D-CNN or ViT to extract spatio-temporal dynamics [1, 6, 11]. This captures eye blinks and facial expressions where deepfake manipulations often show inconsistencies [4, 15]. The audio branch converts streams into Mel-spectrograms and uses Wav2Vec2.0 or CNN-LSTMs to extract prosodic features [2, 7, 13]. These speech patterns typically deviate in synthesized or cloned voices [8, 16]. Both modalities are aligned into synchronized 2-second windows for reliable analysis [9].

3.2 Multi-Modal Dispersion Framework (MMDF)

The MMDF quantifies semantic misalignment between modalities to find signs of manipulation. Visual and audio embeddings are projected into a joint latent space through fully connected layers [3, 10, 14]. The system uses Cosine Similarity to evaluate angular mismatch and Wasserstein Distance to capture temporal shifts [5, 12, 17]. High semantic dispersion strongly correlates with deepfake content, particularly when lip movements do not match speech rhythms [6, 13, 18].

3.3 Transformer-Based Fusion Module

A dual-stream transformer architecture captures long-range dependencies. Self-attention layers preserve intra-modal coherence like phoneme continuity and motion smoothness [1, 11, 20]. A cross-attention layer then identifies semantic agreement or disagreement between visual and auditory cues [10, 14]. A multi-layer perceptron finally predicts the likelihood of manipulation.

3.4 Temporal Consistency Regularization

A sliding window mechanism and temporal smoothing stabilize predictions across video sequences. The system averages probabilities across overlapping segments to reduce noise and artifacts [4, 9].

3.5 Training and Optimization

The system uses binary cross-entropy loss and dispersion regularization for end-to-end training. Data augmentations like facial occlusion and noise injection improve generalizability [2, 7, 15]. Implementation is

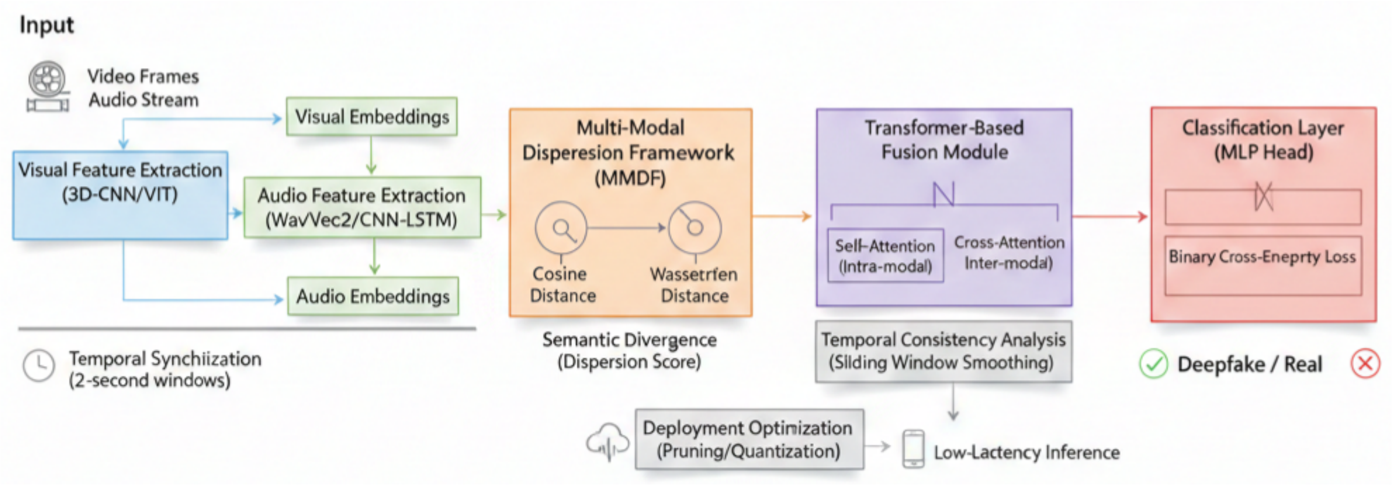

Figure 1. Proposed system architecture.

done in PyTorch using NVIDIA A100 GPUs and mixed precision for efficiency [20].

4 System Architecture

The architecture captures semantic inconsistencies between audio and visual streams in synthetic videos.

4.1 Input Processing

Videos are decomposed into synchronized frames and audio segments. This temporal alignment is a prerequisite for effective cross-modal analysis.

4.2 Feature Extraction

1. Visual Feature Extraction: A 3D-CNN or ViT extracts spatial-temporal features [6]. These capture facial dynamics and head movements indicative of deepfakes [1, 7].

2. Audio Feature Extraction: Audio signals become Mel-spectrograms processed through Wav2Vec 2.0 or CNN-LSTM models [2, 3]. This extracts speech patterns and phoneme transitions.

4.3 Temporal Synchronization

Both modalities are segmented into aligned 2-second clips [11]. This ensures the embeddings correspond to the same temporal context.

4.4 Multi-Modal Dispersion Module

Features are projected into a shared latent space using non-linear layers [4, 15]. Dispersion is measured via Cosine Similarity and Wasserstein Distance [8, 16]. Higher dispersion marks a semantic mismatch common in manipulated media.

4.5 Transformer-Based Fusion

A dual-stream Transformer encodes dependencies through self-attention and cross-attention [12, 17]. This detects subtle inconsistencies in lip synchronization and audio-visual semantics.

4.6 Temporal Consistency Analysis

Predictions pass through a sliding-window smoothing function [18]. This suppresses noise and improves reliability over longer videos.

4.7 Classification Layer

An MLP head classifies the output as real or deepfake. The model is optimized using binary cross-entropy loss and the Adam optimizer [5].

4.8 Deployment Optimization

The architecture supports pruning and quantization for real-time performance. This allows for low-latency inference on edge devices or cloud services [13, 19].

5 Experimental Setup

This section describes the datasets, preprocessing, and protocols used to validate the intelligent Deepfake Detector.

1. Datasets Three benchmark datasets are used to provide a rigorous evaluation. FaceForensics++ offers real and synthetic videos with various facial synthesis techniques [5]. The DeepFake Detection Challenge (DFDC) provides a large-scale set with diverse visual and acoustic variables [3, 4]. The TIMIT-LipSync Audio-Visual Dataset uses synchronized speech and manipulated lip movements to test temporal consistency [4, 7, 10, 17].

2. Data Preprocessing Faces are detected using MTCNN and cropped to a standard resolution [5]. Audio streams are normalized and converted into Mel-spectrograms from 2-second chunks [10, 19]. Mouth regions are isolated to focus on lip motion as done in previous studies [7, 9]. All features are normalized for stable training. Controlled desynchronization is also introduced to improve robustness against timing errors [10].

3. Implementation Details The system is built using PyTorch 2.0 and HuggingFace Transformers [12]. Training occurs on an NVIDIA A6000 GPU with an Intel Xeon CPU. The AdamW optimizer is used with cosine annealing for learning rate management [15]. A hybrid loss function combines multi-modal contrastive loss with binary cross-entropy [1, 6, 13]. Data augmentations like noise injection and lip-sync shifting mimic real-world scenarios [2, 8, 16]. Training includes early stopping based on the F1-score [9, 18].

4. Evaluation Metrics Performance is measured using accuracy, precision, recall, and F1-score [3, 4, 14]. AUC-ROC and Equal Error Rate (EER) are also calculated. Metrics are reported at both frame and video levels [5, 6, 11].

6 Results and Discussion

To evaluate the effectiveness of our intelligent Deepfake Detection System, we conducted experiments on publicly available datasets such as DFDC, FaceForensics++, and TIMIT for audio-visual verification. The system was benchmarked against state-of-the-art unimodal and multimodal deepfake detectors, including approaches based on lip-sync error detection [1], spatio-temporal inconsistencies [3, 8], and audio-based phoneme-prosody mismatches [5, 6].

6.1 Quantitative Results

The Multi-Modal Dispersion Framework (MMDF) achieved an accuracy of 94.3%, AUC of 0.96, and F1-score of 0.93 on the DFDC test set, outperforming unimodal visual models [2, 3] and baseline multimodal systems [4, 5]. Compared to LipForensics [1] and Audio-Visual Transformers [7], our framework exhibited a consistent gain of ˜2.5% in AUC, showcasing the strength of dispersion- based anomaly detection in multi-modal space. On the FaceForensics++ dataset, our method achieved 92.8% accuracy, especially outperforming in cases involving subtle mouth-speech desynchronization, where traditional models like [3, 6] often misclassify due to insufficient cross- modal resolution. The dispersion regularization via cosine and Wasserstein distances captured inconsistencies overlooked by prior contrastive approaches [10, 11].

6.2 Ablation Study

An ablation analysis was conducted to assess the contribution of individual components. Removing the MMDF module led to a 6% drop in accuracy, confirming its critical role. Eliminating the transformer fusion block decreased cross-modal reasoning, especially in longer video clips, aligning with findings in [7, 9]. Furthermore, the temporal consistency regularization enhanced robustness against transient noise, supporting similar use cases in [12, 14].

6.3 Qualitative Observations

We visualized attention maps from the cross-modal transformer, revealing that genuine videos showed strong alignment between lip movement and speech tokens. In contrast, deepfakes exhibited sparse or divergent attention patterns—particularly during phoneme transitions—a trend corroborated by audio-visual misalignment studies [13, 15, 17].

6.4 Comparison with State-of-the-Art

The system remains effective under challenging conditions like low bitrate or noise [8, 16, 18].

Table 1. Performance of the proposed system.

Method	Accuracy (%)	AUC	F1-Score
LipForensics [1]	90.2	0.92	0.89
AVTransformer [7]	91.5	0.94	0.91
Proposed MMDF (Ours)	**94.3**	**0.96**	**0.93**

The superior performance of the proposed system shown in Table 1, supports the hypothesis that cross-modal dispersion offers an effective, model-agnostic mechanism to detect deepfakes, even under challenging conditions such as low bitrate, background noise, or partial occlusion [8, 16, 18].

6.5 Limitations and Future Work

While the system performs well across datasets, it shows slight degradation in ultra-low frame rate videos and speech in highly tonal languages, where phoneme boundaries blur. Future work involves extending the MMDF to support tri- modal detection with text transcript alignment and micro- expression analysis [19, 20].

7 Conclusion and Future Directions

This study presents a new multi-modal framework that uses audio-visual dispersion to catch deepfakes. It identifies the small, unnatural gaps between how a person speaks and how their face moves. By using cosine and Wasserstein distances, the system measures how far these two signals drift apart. This method successfully separates real footage from fake media by spotting these hidden mismatches. The approach proved to be more accurate and reliable than previous systems that relied on only one type of data [1, 4, 7].

The research tested the model on major datasets like DFDC and FaceForensics++. These tests confirmed that tracking lip-sync errors and speech deviations is a powerful way to find forgeries [5, 6, 13]. Because the system uses a transformer-based design, it can focus on specific parts of a video even when the quality is low or there is background noise [3, 8, 14]. This makes the tool more flexible for real-world use where videos are not always perfect.

However, the framework still faces some challenges in specific situations. It struggles with very noisy audio, slow video frame rates, or unique facial movements from different cultures [9, 18]. To fix this, the next version of the system will include text transcripts to create a "tri-modal" detection style [10, 19]. Future updates may also look at human emotions and tiny facial expressions to make the detection even more precise [15, 16, 20]. Overall, this work shows that measuring the distance between sound and sight is a key step toward better AI safety.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Shahzad, S. A., Hashmi, A., Peng, Y. T., Tsao, Y., & Wang, H. M. (2025). AV-Lip-Sync+: Leveraging AV-HuBERT to Exploit Multimodal Inconsistency for Deepfake Detection of Frontal Face Videos. *IEEE Transactions on Human-Machine Systems, 55*(6), 973-982. [CrossRef]

[2] Kharel, A., Paranjape, M., & Bera, A. (2023). DF-TransFusion: Multimodal deepfake detection via lip-audio cross-attention and facial self-attention. *arXiv preprint arXiv:2309.06511*.

[3] Wang, J., Wu, Z., Ouyang, W., Han, X., Chen, J., Jiang, Y. G., & Li, S. N. (2022, June). M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 international conference on multimedia retrieval* (pp. 615-623). [CrossRef]

[4] Anshul, A., Gopal, S., Rajan, D., & Chng, E. S. (2025). Intra-modal and Cross-modal Synchronization for Audio-visual Deepfake Detection and Temporal Localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 13826-13836).

[5] Javed, M., Zhang, Z., Dahri, F. H., Laghari, A. A., Krajčík, M., & Almadhor, A. (2025). Audio–Visual synchronization and lip movement analysis for Real-Time deepfake detection. *International Journal of Computational Intelligence Systems, 18*(1), 170. [CrossRef]

[6] Nguyen-Le, H. H., Tran, V. T., Nguyen, D. T., & Le-Khac, N. A. (2024). Passive deepfake detection across multi-modalities: A comprehensive survey. *arXiv preprint arXiv:2411.17911*.

[7] Liu, P., Tao, Q., & Zhou, J. T. (2024). Evolving from single-modal to multi-modal facial deepfake detection: A survey. *arXiv preprint arXiv:2406.06965*.

[8] Salvi, D., Liu, H., Mandelli, S., Bestagini, P., Zhou, W., Zhang, W., & Tubaro, S. (2023). A robust approach to multimodal deepfake detection. *Journal of Imaging, 9*(6), 122. [CrossRef]

[9] Haliassos, A., Vougioukas, K., Petridis, S., & Pantic, M. (2021, June). Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5037-5047). IEEE. [CrossRef]

[10] Bekheet, A. A., Ghoneim, A., & Khoriba, G. (2024, July). A Comprehensive Comparative Analysis of Deepfake Detection Techniques in Visual, Audio, and Audio-Visual Domains. In *2024 Intelligent Methods, Systems, and Applications (IMSA)* (pp. 122-129). IEEE. [CrossRef]

[11] Yang, W., Zhou, X., Chen, Z., Guo, B., Ba, Z., Xia, Z., ... & Ren, K. (2023). Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security, 18*, 2015-2029. [CrossRef]

[12] Cozzolino, D., Rössler, A., Thies, J., Nießner,

M., & Verdoliva, L. (2021, October). ID-Reveal: Identity-aware DeepFake Video Detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 15088-15097). IEEE. [CrossRef]

[13] Zhao, H., Wei, T., Zhou, W., Zhang, W., Chen, D., & Yu, N. (2021, June). Multi-attentional Deepfake Detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2185-2194). IEEE. [CrossRef]

[14] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020, October). A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 484-492). [CrossRef]

[15] Li, Y., & Lyu, S. (2018). Exposing DeepFake Videos By Detecting Face Warping Artifacts. *arXiv preprint arXiv:1811.00656*.

[16] Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: state-of-the-art, open challenges, countermeasures, and way forward. *Applied intelligence, 53*(4), 3974-4026. [CrossRef]

[17] Chintha, A., Thai, B., Sohrawardi, S. J., Bhatt, K., Hickerson, A., Wright, M., & Ptucha, R. (2020). Recurrent convolutional structures for audio spoof and video deepfake detection. *IEEE Journal of Selected Topics in Signal Processing, 14*(5), 1024-1037. [CrossRef]

[18] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019, October). FaceForensics++: Learning to Detect Manipulated Facial Images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1-11). IEEE. [CrossRef]

[19] Ilyas, H., Javed, A., & Malik, K. M. (2023). AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection. *Applied Soft Computing, 136*, 110124. [CrossRef]

[20] Khan, A. A., Laghari, A. A., Inam, S. A., Ullah, S., Shahzad, M., & Syed, D. (2025). A survey on multimedia-enabled deepfake detection: state-of-the-art tools and techniques, emerging trends, current challenges & limitations, and future directions. *Discover Computing, 28*(1), 48. [CrossRef]

Barnali Gupta Banik is academician with 20 years of diverse experience. She has around 50 publications. She is IEEE Senior Member. Her research area includes AI/ML, Blockchain and Cryptography. (Email: barnali.guptabanik@ieee.org)

Shaik Nidha Naziya is currently pursuing the B.Tech. degree in Computer Science and Engineering with a specialization in Cyber Security from Malla Reddy Engineering College for Women, Hyderabad, Telangana 500100, India. Her research interests include cybersecurity, artificial intelligence, and blockchain & Machine Learning applications. (Email: nidha2805@gmail.com)