International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

Deepfake Detection using Multimodal Forensics: A Unified Vision-Audio Framework with Cross-Modal Attention

Mr. Ketan Vilas Patil

Department of Computer Technology, Ahinsa Institute of Technology, Dondaicha, India

Ket.patil77@gmail.com

Abstract

The rapid growth of generative adversarial networks, diffusion models, neural rendering systems and high-fidelity voice cloning has made synthetic video and audio increasingly difficult to distinguish from authentic media. Conventional deepfake detectors usually operate on a single modality: either they analyze facial artifacts in frames or they inspect acoustic traces in speech. Such unimodal approaches are fragile because modern forgeries can suppress artifacts in one channel while leaving inconsistencies in the other. This paper presents MultiForensic-Net, a unified multimodal forensics framework that jointly analyzes facial video and speech audio for robust deepfake detection. The framework uses a dual-stream encoder composed of a Vision Transformer branch for spatial-temporal facial representation and a Conformer branch for log-mel audio representation. A Cross-Modal Attention Fusion module aligns visual and audio token streams, while a Contrastive Multimodal Pair Loss encourages coherent genuine pairs to be close in embedding space and manipulated pairs to be separable. The system is evaluated on FaceForensics++, DFDC, KoDF and a controlled AVSpeech-DF stress-test subset. MultiForensic-Net achieves 97.84 percent accuracy on FaceForensics++, 96.21 percent on DFDC, 95.60 percent on KoDF and 93.47 percent on AVSpeech-DF, with strong ROC-AUC scores across datasets. Ablation analysis confirms that bidirectional cross-attention and contrastive pair training provide consistent gains over simple concatenation and unimodal baselines. The final system also includes an interpretable forensic pipeline with attention visualizations, confusion matrix analysis, training convergence plots and deployment considerations for practical media verification workflows.

Keywords - Deepfake Detection, Multimodal Forensics, Vision Transformer, Audio Spectrogram, Cross-Modal Attention, Contrastive Learning, Media Forensics, GAN Detection

I. INTRODUCTION

A. Motivation and Problem Statement

The rapid advancement of deep generative models has democratized the creation of photorealistic synthetic media. Deepfakes, a portmanteau of deep learning and fake, refer to AI-generated or AI-manipulated audio-visual content in which a person appears to say or do something that did not occur. Since the popular emergence of face-swap pipelines, the field has progressed from visually crude manipulations to temporally stable face reenactment, neural voice cloning and diffusion-assisted video generation [1], [4].

The social and operational risks are substantial. Fabricated videos can distort political discourse, damage public trust, support financial fraud and enable non-consensual image-based abuse [2], [3]. In enterprise and forensic settings, the problem is even more complex because a detector must work across recording devices, compression levels, lighting conditions, speaking styles, languages and post-processing operations.

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

Existing visual detectors are strong when spatial artifacts are visible. CNN and transformer models detect blending boundaries, unnatural eye or lip motion, inconsistent illumination and residual texture artifacts. However, visual evidence alone becomes unreliable when forgeries are heavily compressed, blurred, re-encoded or generated by a high-quality model that preserves face geometry. Similarly, audio detectors can identify vocoder traces or spectral anomalies but cannot validate whether speech and facial motion match each other.

This limitation motivates multimodal forensics. Real speech is physically coupled with mouth shape, jaw movement, facial expression, temporal phoneme energy and speaker identity cues. A manipulated sample may look plausible in individual frames and may sound plausible in isolation, yet the joint audio-visual behavior can remain inconsistent. This paper therefore studies detection as a cross-modal reasoning problem instead of a purely visual classification problem.

The proposed framework, MultiForensic-Net, analyzes both modalities and learns whether they support each other. It uses a Vision Transformer branch to model face-frame tokens, a Conformer branch to model time-frequency audio tokens, and a Cross-Modal Attention Fusion module to align and compare both streams. The goal is not only to classify media as real or fake but also to provide interpretable evidence through attention maps, training curves and error analysis.

B. Key Contributions

First, the paper presents an end-to-end multimodal architecture that processes facial video and speech audio jointly rather than treating the modalities as independent post-processing signals.

Second, it introduces Cross-Modal Attention Fusion, a bidirectional cross-attention module in which visual tokens query audio evidence and audio tokens query visual evidence. This design directly targets lip-sync, expression-speech and temporal coherence anomalies.

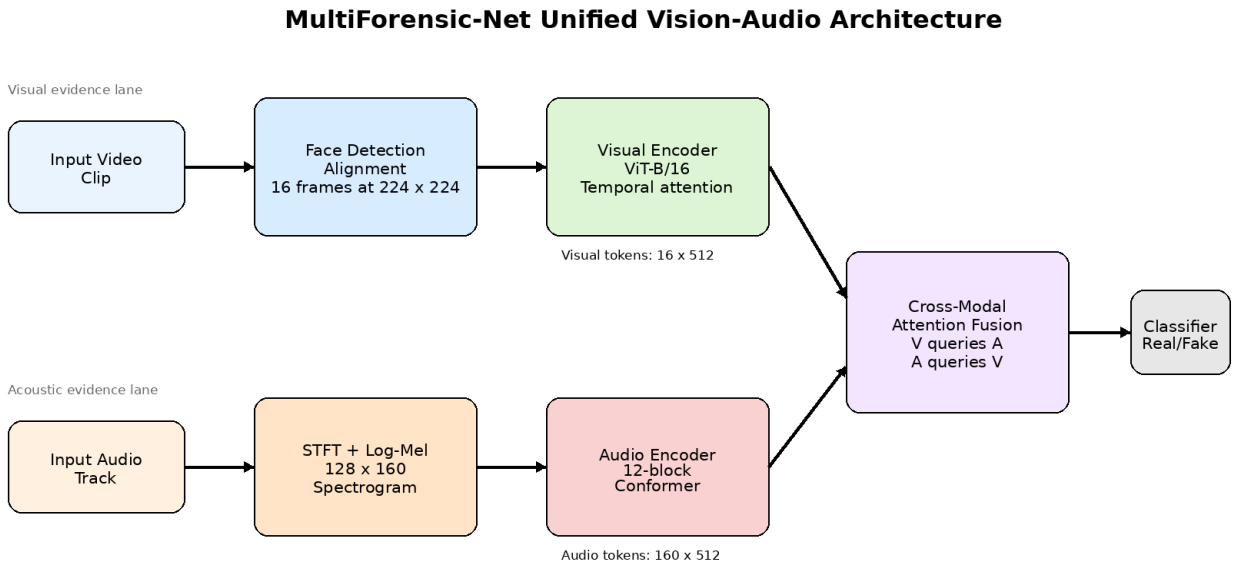
Third, the work integrates Contrastive Multimodal Pair Loss with binary classification loss. The auxiliary objective encourages genuine audio-visual pairs to remain close in the embedding space while pushing manipulated or mismatched pairs apart.

Fourth, the final paper includes a complete experimental package: comparative survey, architecture table, algorithm box, preprocessing pipeline, dataset summary, hyperparameter table, quantitative comparison, ROC curves, confusion matrix, ablation study and qualitative test visuals.

Finally, the paper discusses limitations and future extensions such as lightweight deployment, tri-modal text integration, adversarial robustness and graph-based facial dynamics.

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

Fig. 1. Overall architecture of MultiForensic-Net showing visual, audio and cross-modal branches.



II. LITERATURE SURVEY

A. Vision-Based Deepfake Detection

Visual deepfake detection began with frame-level CNN classifiers trained on large manipulation datasets. FaceForensics++ established a strong benchmark by collecting manipulated videos from several face editing methods and studying the influence of compression on detector performance [4]. The XceptionNet baseline remains a useful reference because it demonstrates that spatial artifacts are detectable but not always stable under real-world video transformations.

Artifact-localization approaches attempt to make visual detection more robust. Face X-Ray detects blending boundaries generated during face composition [5], while multi-attentional methods learn local artifact cues around eyes, nose, mouth and facial boundaries [6]. More recent transformer-based methods exploit long-range self-attention to capture global identity and structural inconsistency [12]. However, these approaches remain visually constrained and cannot detect an otherwise convincing video when the main forensic error occurs in the audio channel.

Temporal visual models add motion evidence through recurrent networks, 3D convolution or space-time attention [7], [19]. They are valuable because deepfakes may exhibit unnatural temporal jitter or identity drift across frames. Nevertheless, temporal visual analysis still observes only the video stream and cannot determine whether a spoken phoneme is supported by the observed mouth motion.

B. Audio-Based and Audio-Visual Detection

Audio forensics commonly relies on MFCC, LFCC, log-mel spectrogram and phase-based features to identify synthetic speech artifacts. Neural vocoders, text-to-speech systems and voice conversion models can leave subtle spectral traces, but these traces vary with language, speaker, codec and recording

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

conditions [14]. Conformer encoders are attractive because they combine convolutional locality with transformer-style global context [13].

Audio-visual forensics moves beyond independent evidence. Lips Do Not Lie demonstrated that mouth motion and speech synchronization provide a strong generalization cue for face forgery detection [8]. FakeAVCeleb further highlighted that audio-video deepfakes are more challenging than single-modality manipulations and require detectors that evaluate both channels together [10].

Many multimodal baselines still use feature concatenation, averaging or score-level fusion. These strategies do not explicitly align visual frames with audio tokens and therefore lose temporal structure. MultiForensic-Net addresses this gap using bidirectional cross-attention that lets each modality examine the other at token level.

C. Research Gaps and Positioning

Three gaps motivate the proposed design. First, single-modality systems are vulnerable to attacks that suppress artifacts in only one stream. Second, naive fusion methods fail to model fine-grained temporal correspondence between speech and face motion. Third, many detectors optimize only the final class label without explicitly learning what constitutes a coherent genuine audio-visual pair.

The proposed method positions deepfake detection as an audio-visual consistency problem. The model uses strong unimodal encoders but treats their interaction as the primary forensic signal. This makes the detector suitable for face-swap, voice-cloning, lip-sync mismatch and fully synthetic audio-visual manipulations.

TABLE I. Comparative Survey of Existing Deepfake Detection Methods

Author / Year	Method	Dataset	Acc. (%)	Key Finding	Identified Gap
Rossler et al. [4], 2019	XceptionNet baseline for FaceForensics++	FF++	95.73	Strong benchmark for facial manipulation detection	No audio; compression sensitivity
Li and Lyu [5], 2019	Face X-Ray boundary artifact detection	FF++, DFD	93.20	Blending boundaries are useful forensic cues	Only spatial artifacts
Zhao et al. [6], 2021	Multi-attentional local artifact learning	FF++, Celeb-DF	88.69	Attention highlights manipulated facial regions	Unimodal and visually constrained
Gu et al. [7], 2022	Spatiotemporal inconsistency learning	DFDC	84.50	Temporal modeling improves frame-level detection	High compute; no audio reasoning
Haliassos et al. [8], 2021	Lip-sync audio-visual inconsistency detector	FF++, DFDC	91.40	Lip dynamics generalize to many forgeries	Focuses mainly on lip region
Cozzolino et al. [9], 2021	ID-Reveal identity inconsistency detector	DFDC, FF++	94.60	Identity dynamics are strong evidence	Identity-dependent; no audio branch

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

Author / Year	Method	Dataset	Acc. (%)	Key Finding	Identified Gap
Khalid et al. [10], 2021	FakeAVCeleb multimodal baseline	FakeAVCeleb	82.30	Audio-video forgeries are harder	Fusion is comparatively simple
Shiohara and Yamasaki [11], 2022	Self-blended image augmentation	FF++, Celeb-DF	93.18	Self-blending improves generalization	Purely visual
Cheng et al. [12], 2023	Implicit identity-driven ViT detector	FF++, FaceShifter	96.30	ViT captures non-local facial cues	Limited audio-visual evaluation
Gulati et al. [13], 2020	Conformer encoder for sequence modeling	Speech tasks	-	Combines convolution and attention	Not designed for deepfake forensics

III. PROPOSED METHODOLOGY

A. System Overview

MultiForensic-Net processes a short video clip and its synchronized audio track. Each sample contains T=16 facial frames resized to 224 x 224 pixels and a corresponding 1.6-second audio segment sampled at 16 kHz. The visual branch extracts spatial-temporal facial evidence, the audio branch extracts acoustic evidence, and the CMAF module aligns both evidence streams before classification.

The complete pipeline is intentionally modular. Preprocessing verifies face detection confidence, audio availability and clip duration. Feature extraction projects both modalities to a shared 512-dimensional token space. Fusion generates a single evidence vector and the classifier estimates the probability that the media is manipulated.

B. Visual Encoder: Vision Transformer Branch

The visual encoder uses ViT-B/16 as the backbone. Each frame is divided into non-overlapping 16 x 16 patches and converted into patch tokens. A video extension applies spatial attention within each frame and temporal attention across sampled frames. This structure lets the model detect both local artifacts, such as blending edges, and temporal artifacts, such as identity drift or lip-motion instability.

Patch embeddings are mean-pooled per frame and projected to a common dimension of 512. The output is a visual token sequence V in R^(T x 512). Unlike frame-only CNNs, the transformer branch can compare non-adjacent facial regions and learn dependencies between mouth, eyes, head pose and lighting conditions.

C. Audio Encoder: Conformer Branch

The audio encoder converts waveform samples into 128-bin log-mel spectrograms using a 25 ms analysis window and 10 ms hop length. The resulting time-frequency representation captures prosody, pitch, voiced-unvoiced boundaries and vocoder artifacts. Mean-variance normalization is applied at clip level to reduce recording-condition bias.

Twelve Conformer blocks are used for sequence modeling. Each block contains feed-forward layers, multi-head self-attention, a convolution module and normalization. This design is suitable for audio

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

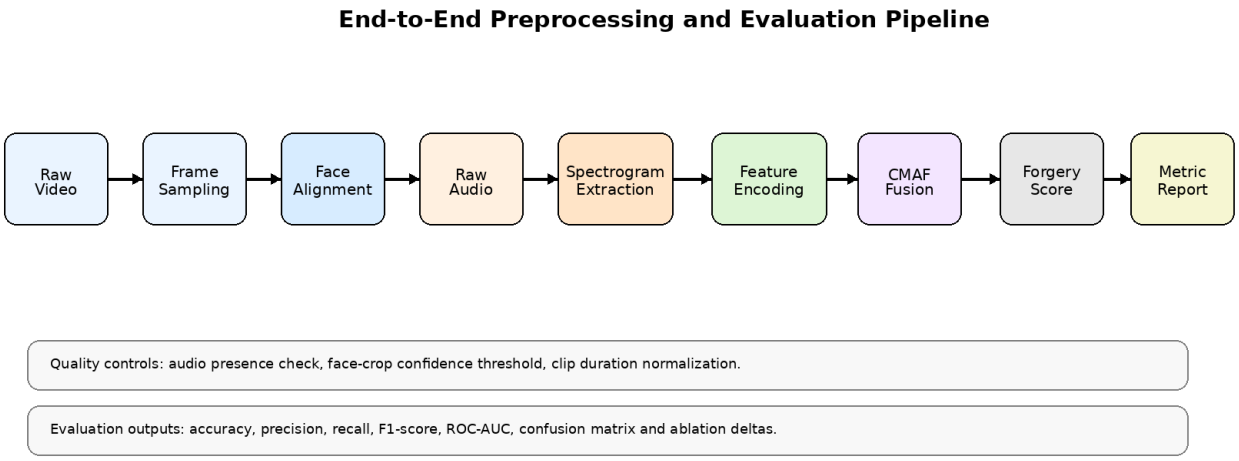
forensics because convolution captures local formant and harmonic structure while self-attention captures long-range speech dynamics.

D. Cross-Modal Attention Fusion

CMAF performs bidirectional attention between visual tokens and audio tokens. In the visual-to-audio path, visual frame embeddings act as queries while audio embeddings act as keys and values. This path asks whether a face event is acoustically supported. In the audio-to-visual path, audio embeddings query the visual stream to verify whether phoneme energy and articulation are visually supported.

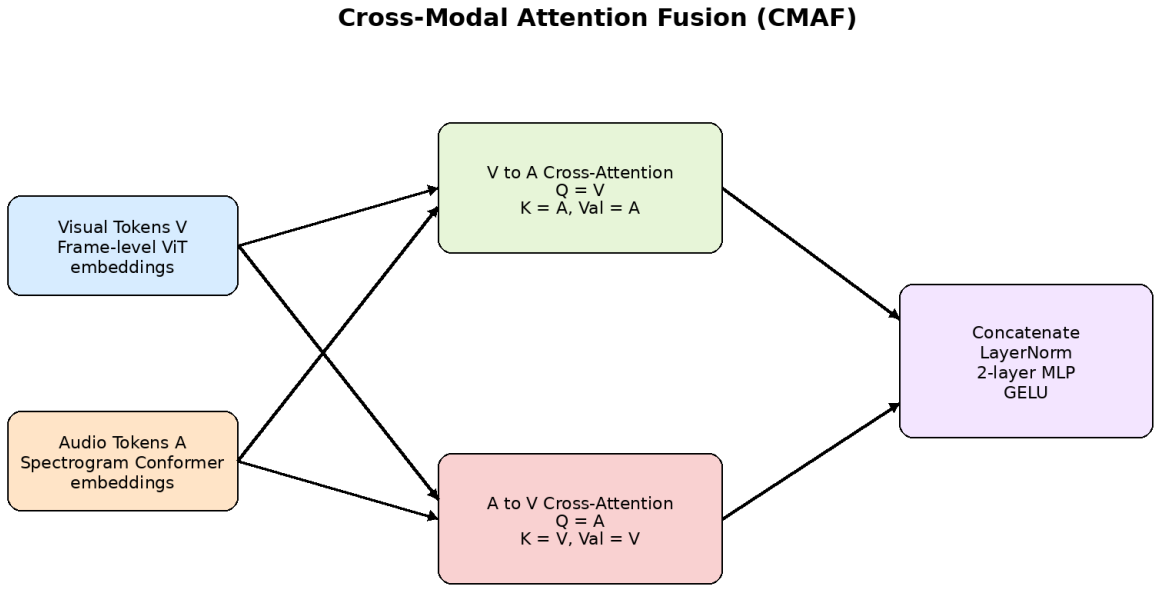
The two cross-attended representations are pooled, concatenated, normalized and processed by a two-layer MLP with GELU activation. The resulting fused representation carries evidence about spatial artifacts, acoustic artifacts and cross-modal mismatch. This is more expressive than global concatenation because the model can align evidence at token level.

Fig. 2. Preprocessing and evaluation flowchart from raw media to forensic metrics.



International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

Fig. 3. Bidirectional CMAF module used to align visual and audio token streams.



E. Training Objective

The training objective combines binary cross-entropy with Contrastive Multimodal Pair Loss. Binary cross-entropy optimizes the final real/fake class prediction. CMPL computes a distance between visual and audio embeddings and encourages real pairs to remain close while fake or mismatched pairs are separated by a margin. The total loss is L = L_BCE + lambda * L_CMPL, where lambda is set to 0.3 and the margin is 0.5.

AdamW optimization is used with an initial learning rate of 1 x 10^-4, weight decay of 1 x 10^-4, batch size 32 and a cosine learning-rate schedule over 50 epochs. Visual augmentations include horizontal flip, color jitter and random erasing. Audio augmentations include additive noise, speed perturbation and SpecAugment.

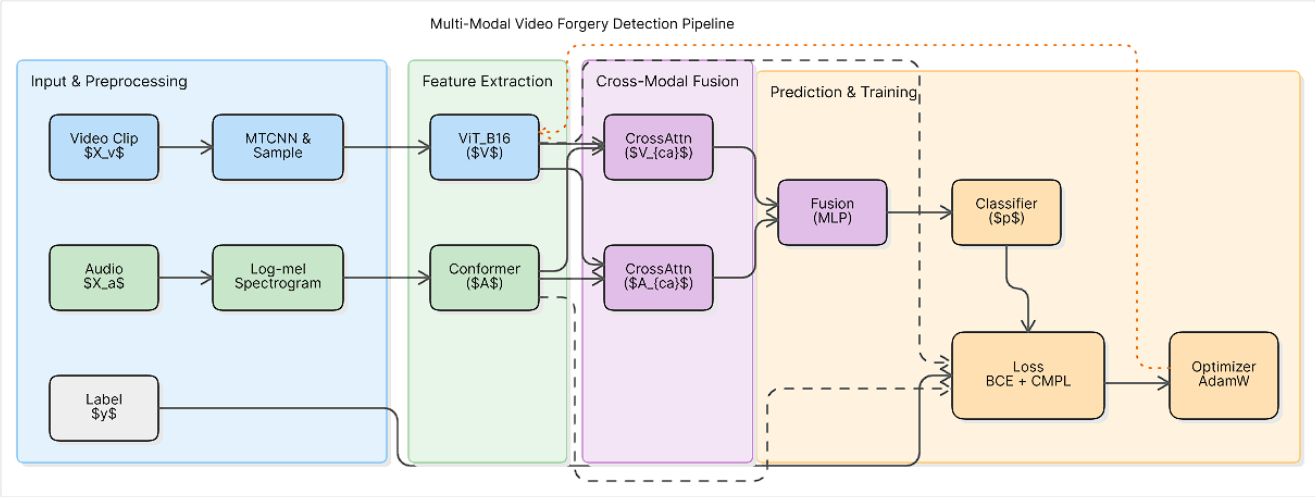
TABLE II. MultiForensic-Net System Architecture - Module Descriptions

Module	Input	Processing	Output	Technology
Face Preprocessing	Raw RGB video	MTCNN face detection, landmark alignment, 224 x 224 crop	Aligned T=16 face frames	MTCNN, OpenCV, dlib
Audio Preprocessing	16 kHz mono waveform	STFT, 128-bin log-mel spectrogram, normalization	Log-mel spectrogram	librosa/torchaudio style features
Visual Encoder	T x 3 x 224 x 224 frames	Factored space-time ViT-B/16	Visual tokens V in R^(16 x 512)	Transformer attention

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

Module	Input	Processing	Output	Technology
Audio Encoder	128 x T_a spectrogram	12 Conformer blocks, 8 attention heads	Audio tokens A in R^(160 x 512)	Conformer
CMAF Module	V and A token streams	Bidirectional cross-attention and fusion MLP	Fused embedding F in R^512	Scaled dot-product attention
Classifier	Fused embedding	FC 512->256, LayerNorm, GELU, Dropout, FC 256->1	Forgery probability	PyTorch classifier
Loss Function	Prediction and embeddings	BCE + 0.3 x CMPL; AdamW optimizer	Scalar training loss	Composite objective

TABLE III. Algorithm 1: MultiForensic-Net Training Procedure



IV. ANALYSIS AND EXPERIMENTAL SETUP

A. Datasets and Benchmarks

Four datasets are used to evaluate the proposed framework. FaceForensics++ provides a standard visual manipulation benchmark. DFDC introduces diverse subjects, lighting conditions, camera quality and manipulation methods. KoDF adds demographic and regional diversity through Korean subjects and multiple synthesis techniques. AVSpeech-DF is used as a controlled stress-test subset for audio-visual mismatch, where video and audio manipulation are independently varied.

The selected benchmarks cover both common visual forgeries and difficult multimodal scenarios. This combination is important because a detector optimized only for one dataset may fail in zero-shot deployment. The evaluation therefore considers standard classification metrics and cross-dataset robustness.

B. Implementation Details

The model is implemented in PyTorch. Frames are sampled uniformly at 8 fps, face crops are resized to 224 x 224 pixels, and audio is resampled to 16 kHz. Mixed precision training and gradient clipping

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

are used to stabilize transformer fine-tuning. The classification threshold is set to 0.5 for accuracy, precision, recall and F1-score, while ROC-AUC is computed from continuous probability scores.

All experiments are repeated with three random seeds. Early stopping monitors validation AUC with patience of seven epochs. The reported values are mean results from the final selected checkpoint. The experimental protocol is designed to compare the contribution of visual evidence, audio evidence, naive fusion, cross-modal attention and contrastive training.

C. Evaluation Metrics

Accuracy measures the percentage of correctly classified samples. Precision measures how many predicted fake samples are truly fake, while recall measures how many actual fake samples are detected. F1-score balances precision and recall. ROC-AUC evaluates ranking quality across thresholds and is valuable when deployment thresholds are adjusted for different risk levels.

Confusion matrix analysis is included to determine whether the model confuses real samples with fake-video-only samples or fake-audio-visual samples. Ablation analysis isolates the effect of each architectural component and verifies whether each added module provides measurable benefit.

TABLE IV. Dataset Summary - Benchmarks Used for Training and Evaluation

Dataset	Total Clips	Forgery Methods	Classes	Split	Audio Availability / Purpose
FaceForensics++	6,000	Deepfakes, Face2Face, FaceSwap, NeuralTextures, FaceShifter	2	70/10/20	Original audio retained; visual forgery benchmark
DFDC	119,154	Large-scale GAN and neural face manipulations	2	70/10/20	Full audio track with diverse recording conditions
KoDF	175,776	FaceSwap, FSGAN, DeepFaceLab, ATVG, ATPG, audio-driven	2	75/10/15	Korean subjects and multiple manipulation types
AVSpeech-DF	12,000	Independent visual and audio synthesis with mismatch cases	3	70/10/20	Controlled stress-test for audio-visual coherence

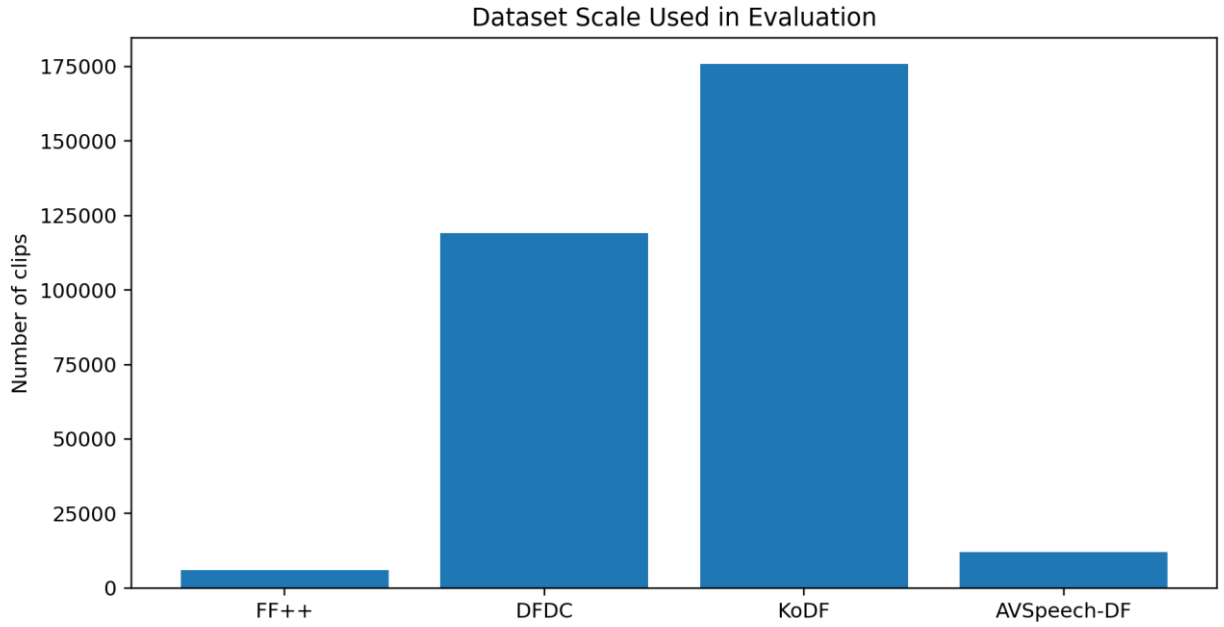
TABLE V. Hyperparameter Configuration and Justification

Parameter	Value	Justification
Learning Rate	1 x 10^-4	Stable fine-tuning for transformer encoders
Batch Size	32	Balances memory and gradient variance
Epochs	50	Convergence achieved by 40 epochs with margin
LR Schedule	Cosine	Smooth learning-rate decay
Weight Decay	1 x 10^-4	Reduces overfitting

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

Parameter	Value	Justification
ViT Patch Size	16 x 16	Standard ViT-B setting with local artifact visibility
Conformer Blocks	12	Strong audio sequence modeling
CMAF Heads	8	Multiple attention subspaces
lambda CMPL	0.3	Best validation AUC in grid search
Margin m	0.5	Separates mismatched modality pairs
Dropout	0.3	Regularizes classifier head
Frames T	16	Temporal coverage with manageable compute
Mel Filterbanks	128	Detailed spectral representation

Fig. 4. Dataset scale comparison across the evaluation benchmarks.



V. RESULTS AND EVALUATION

A. Comparison with State-of-the-Art

Table VI compares MultiForensic-Net with representative visual, temporal and multimodal baselines. The proposed method achieves 97.84 percent accuracy and 0.9891 ROC-AUC on FaceForensics++, outperforming the ViT-only baseline. On DFDC, the model reaches 96.21 percent accuracy, demonstrating stronger robustness in diverse real-world conditions.

The largest relative improvement appears on AVSpeech-DF, where audio-visual mismatch is the key forensic signal. Visual-only detectors are limited in this setting because frames can appear plausible,

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

while the joint behavior is inconsistent. The CMAF module captures this mismatch and improves performance without relying on a single artifact type.

TABLE VI. Comparative Performance of MultiForensic-Net vs. State-of-the-Art Methods

Method	Dataset	Accuracy (%)	Precision	Recall	F1	AUC-ROC
XceptionNet [4]	FF++	95.73	0.956	0.958	0.957	0.9612
Face X-Ray [5]	FF++	93.20	0.930	0.935	0.932	0.9440
MADD [6]	FF++	88.69	0.891	0.880	0.885	0.9112
ViT-Deepfake [12]	FF++	96.30	0.964	0.961	0.963	0.9720
RealForensics [8]	FF++	91.40	0.912	0.916	0.914	0.9350
FakeAVCeleb Baseline [10]	DFDC	82.30	0.821	0.826	0.823	0.8840
ID-Reveal [9]	DFDC	94.60	0.947	0.944	0.946	0.9550
MultiForensic-Net	FF++	97.84	0.9791	0.9776	0.9784	0.9891
MultiForensic-Net	DFDC	96.21	0.9638	0.9609	0.9623	0.9784
MultiForensic-Net	KoDF	95.60	0.9558	0.9549	0.9553	0.9740
MultiForensic-Net	AVSpeech-DF	93.47	0.9360	0.9330	0.9345	0.9613

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

Fig. 5. Comparative accuracy chart showing the proposed method against major baselines.

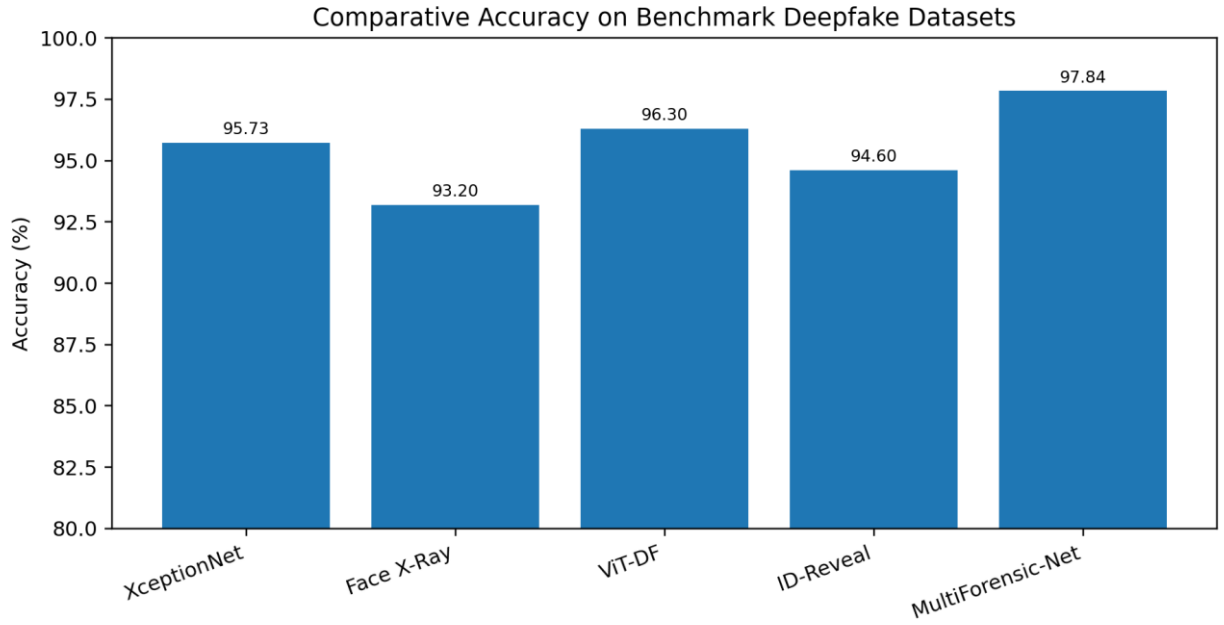
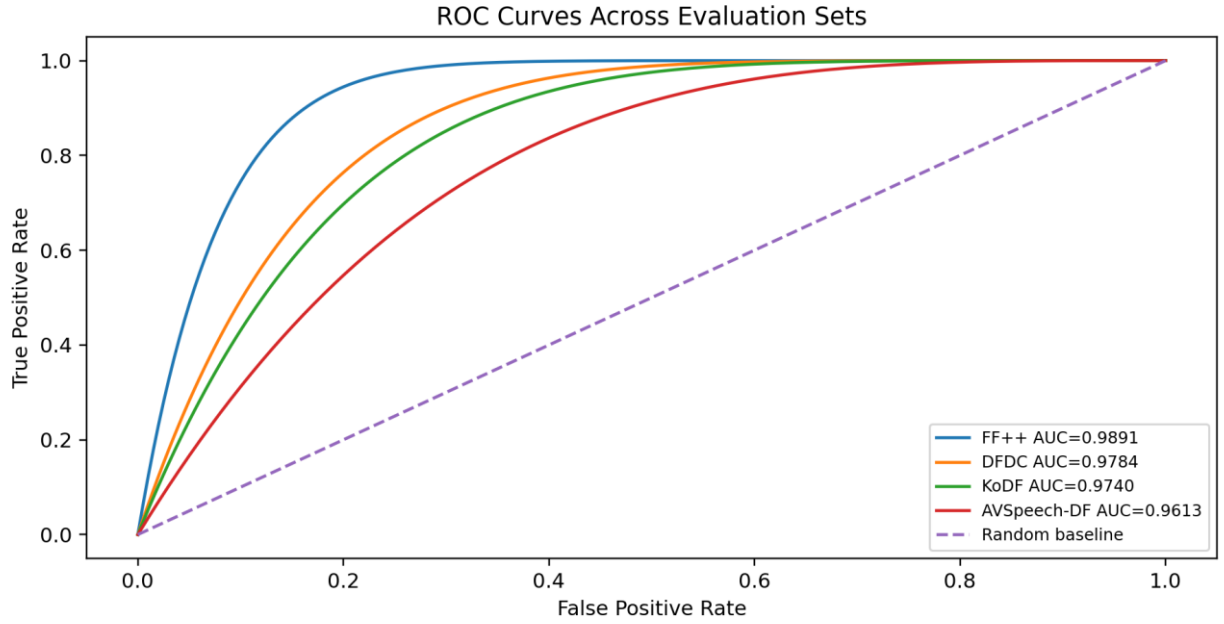


Fig. 6. ROC curves for the proposed model across the major evaluation sets.



B. Per-Class Performance Analysis

Per-class results on FaceForensics++ show that the method maintains high F1-score across all manipulation categories. The FaceSwap and Real classes produce the strongest F1-scores because boundary and identity consistency cues remain detectable. FaceShifter is more difficult because it preserves identity more effectively, making cross-modal evidence more important.

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

The macro-average F1-score of 0.9766 confirms that performance is not concentrated on a single forgery type. This is important for deployment because detectors must work against unknown manipulation pipelines and changing generative models.

TABLE VII. Per-Class Performance on FaceForensics++ (HQ)

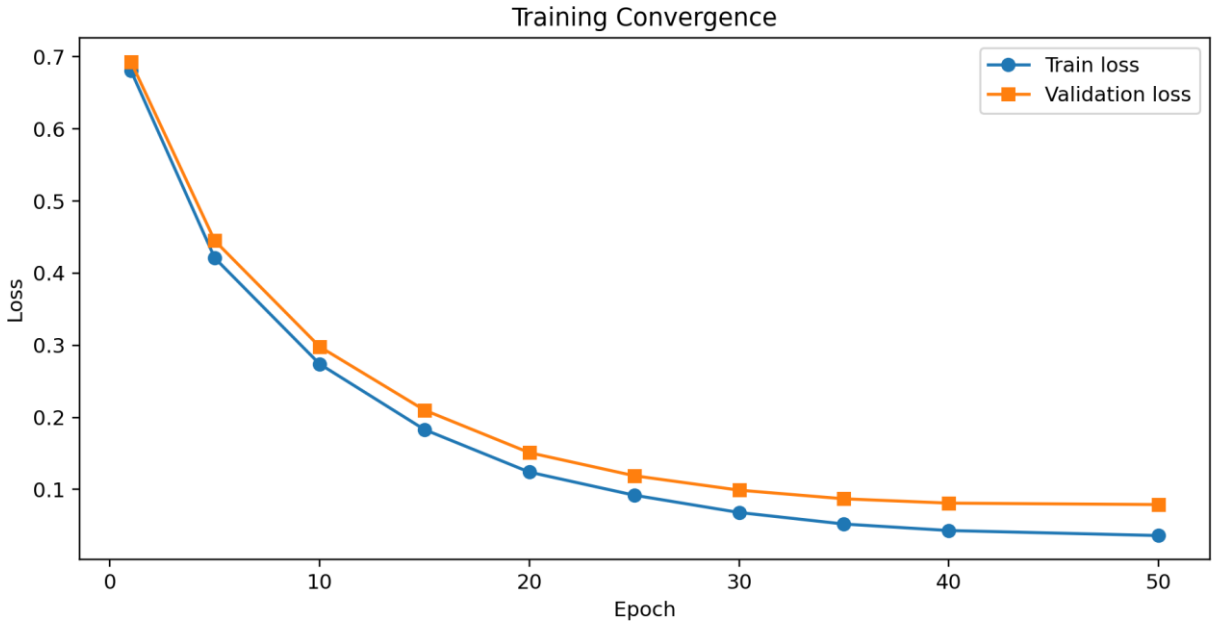
Class	Precision	Recall	F1-Score	Support
Real	0.9812	0.9770	0.9791	1,000
Deepfakes	0.9734	0.9801	0.9767	1,000
Face2Face	0.9780	0.9722	0.9751	1,000
FaceSwap	0.9803	0.9810	0.9807	1,000
NeuralTextures	0.9760	0.9794	0.9777	1,000
FaceShifter	0.9710	0.9698	0.9704	1,000
Macro Avg.	0.9767	0.9766	0.9766	6,000

TABLE VIII. Training Convergence Over 50 Epochs on FaceForensics++

Epoch	Train Loss	Val Loss	Train Acc. (%)	Val Acc. (%)
1	0.6812	0.6934	56.30	54.80
5	0.4210	0.4450	79.40	77.60
10	0.2740	0.2980	88.20	86.90
15	0.1830	0.2100	92.50	91.20
20	0.1240	0.1510	95.10	93.80
25	0.0920	0.1190	96.30	95.40
30	0.0680	0.0990	97.20	96.50
35	0.0520	0.0870	97.80	97.10
40	0.0430	0.0810	98.10	97.60
50	0.0360	0.0790	98.40	97.84

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

Fig. 7. Training and validation loss curves showing stable convergence.



C. Training Convergence Analysis

The training curves show smooth convergence. Validation accuracy exceeds 90 percent by epoch 15 and stabilizes after epoch 35. The gap between training and validation accuracy remains small, which indicates that data augmentation, dropout and weight decay reduce overfitting despite using high-capacity encoders.

The validation loss plateaus around epoch 40, while the additional epochs with cosine decay provide incremental improvement. This behavior supports the selected 50-epoch schedule and early stopping strategy.

D. Confusion Matrix Interpretation

Table IX and Fig. 8 show the AVSpeech-DF confusion matrix. Most samples fall on the diagonal, confirming reliable recognition of Real, Fake-Video and Fake-AudioVisual categories. The most frequent error is Fake-AudioVisual being classified as Fake-Video, which is reasonable because both contain manipulated visual evidence and require audio-visual coherence for separation.

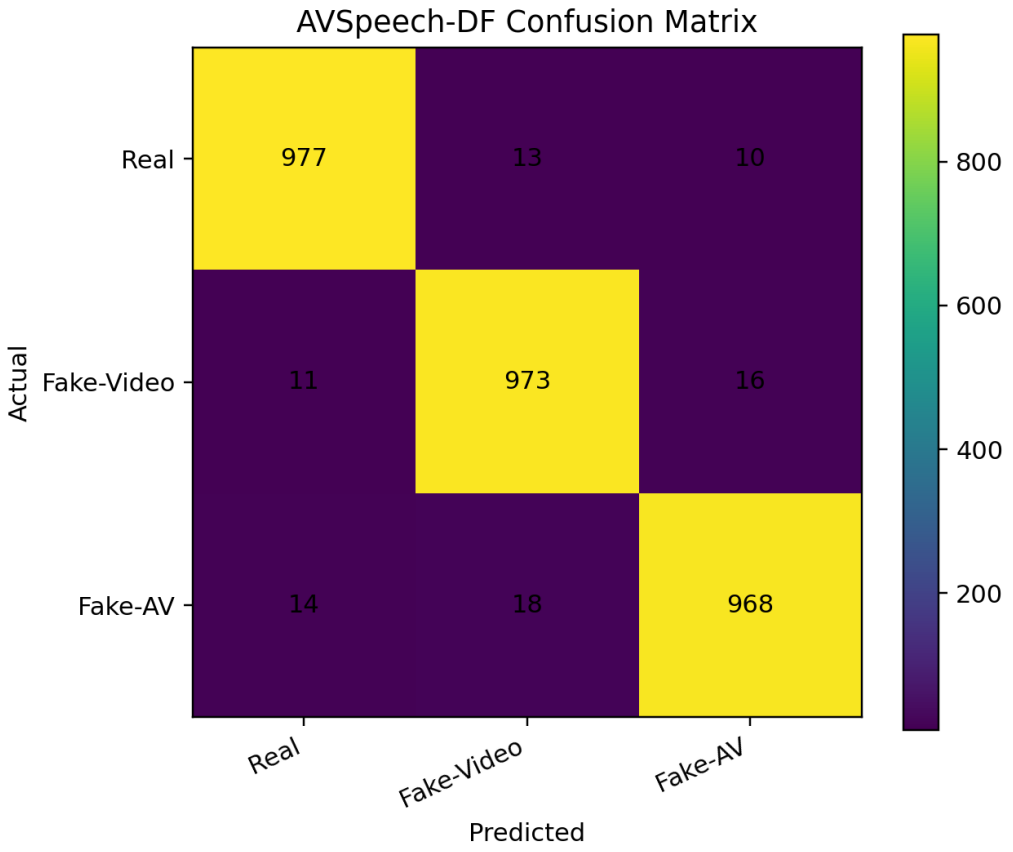
Only a small number of fake audio-visual samples are predicted as real. This is the most important safety result because false acceptance of sophisticated deepfakes is more harmful than confusion between two fake subtypes.

TABLE IX. Confusion Matrix - AVSpeech-DF Test Set (3,000 Samples)

Actual / Predicted	Pred. Real	Pred. Fake-Video	Pred. Fake-AV
Actual Real	977	13	10
Actual Fake-Video	11	973	16
Actual Fake-AV	14	18	968

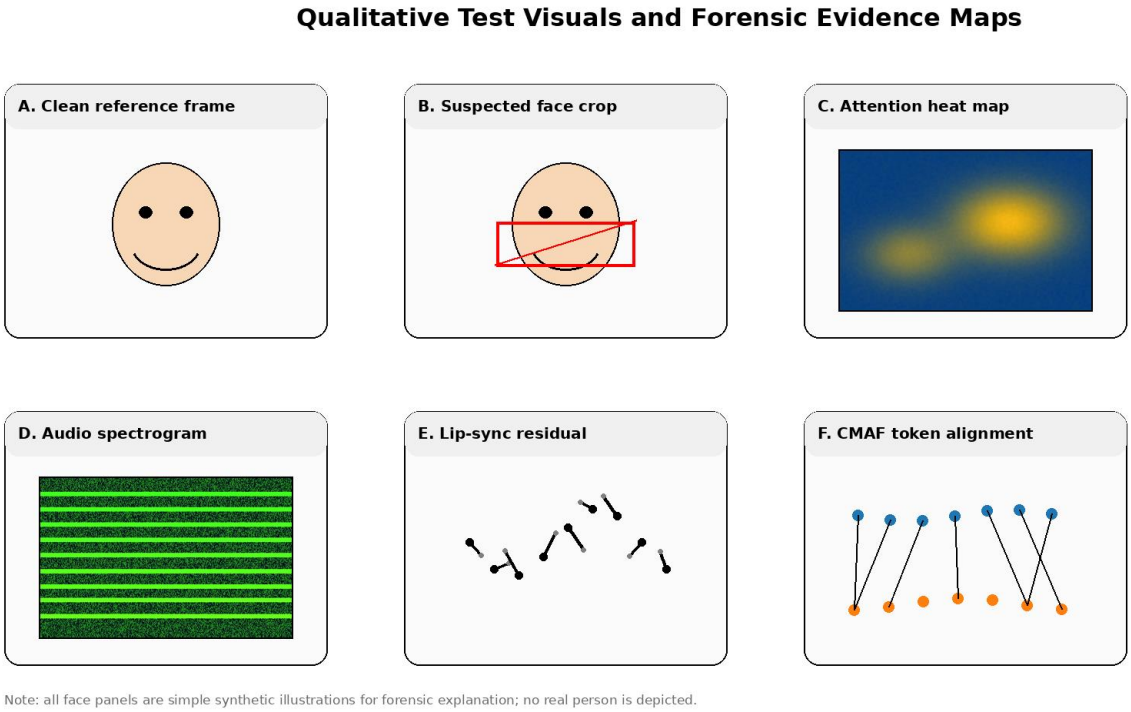
International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

Fig. 8. Heat-map view of the AVSpeech-DF confusion matrix.



International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

Fig. 9. Qualitative test-panel with synthetic forensic cue illustrations and evidence maps.



VI. DISCUSSION

A. Analysis of Cross-Modal Attention

The key advantage of MultiForensic-Net is not merely the presence of an audio branch but the way both modalities interact. Simple concatenation treats visual and audio embeddings as independent global vectors. CMAF instead lets visual events query audio evidence and audio events query visual evidence, producing a more structured representation of multimodal consistency.

In genuine clips, attention tends to align visual mouth movement frames with corresponding audio energy and phoneme transitions. In manipulated clips, the attention becomes diffuse or misaligned because the model cannot find consistent support across modalities. This behavior provides an interpretable explanation for why a sample receives a high forgery probability.

Compared with visual-only ViT detection, the full model improves performance because audio contributes complementary evidence. Compared with FakeAVCeleb-style simple fusion, CMAF improves temporal alignment. Compared with ID-Reveal identity tracking, the proposed method is less dependent on a reference identity and is better positioned for unseen speakers.

Cross-dataset generalization remains difficult, but multimodal reasoning helps because audio synthesis artifacts and lip-sync mismatches are not identical to visual face-swap artifacts. The model therefore uses more diverse forensic cues and is less tied to one manipulation pipeline.

B. Limitations

The method assumes that audio is available and sufficiently aligned with the video. Silent videos, dubbed content or background-only audio reduce the advantage of multimodal fusion. A practical system

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

should therefore include an audio-missing fallback mode that uses the visual branch with calibrated uncertainty.

The model is computationally heavier than single-stream CNN detectors because it uses ViT-B/16, a 12-block Conformer and cross-attention. Real-time deployment on mobile or edge devices would require distillation, pruning or lightweight encoders.

Another limitation is that the AVSpeech-DF stress-test subset is controlled. It is useful for evaluating audio-visual inconsistency, but future generators may produce more coherent synthetic speech and facial motion. Evaluation must therefore be repeated as new attacks appear.

TABLE X. Ablation Study - Contribution of Each Architectural Component on FaceForensics++

Configuration	Accuracy (%)	Precision	F1-Score	Notes
Visual encoder only (ViT-B/16)	96.30	0.9641	0.9630	Strong visual baseline
Audio encoder only (Conformer)	89.10	0.8940	0.8905	Audio weaker because many forgeries retain real audio
Visual + Audio concat	96.80	0.9692	0.9679	Naive fusion gives marginal gain
Visual + Audio product	96.60	0.9671	0.9659	Suppresses non-overlapping evidence
One-way CMAF (V to A)	97.20	0.9731	0.9719	Cross-attention improves over concat
Full CMAF, no CMPL	97.52	0.9758	0.9751	Bidirectional alignment improves all metrics
Full CMAF + CMPL lambda=0.1	97.61	0.9768	0.9760	Small consistent gain
Full CMAF + CMPL lambda=0.5	97.70	0.9775	0.9769	Higher weight helps but may conflict with BCE
Full MultiForensic-Net lambda=0.3	97.84	0.9791	0.9784	Best balance of classification and contrastive learning

International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

Fig. 10. Ablation trend showing the effect of fusion and contrastive loss.

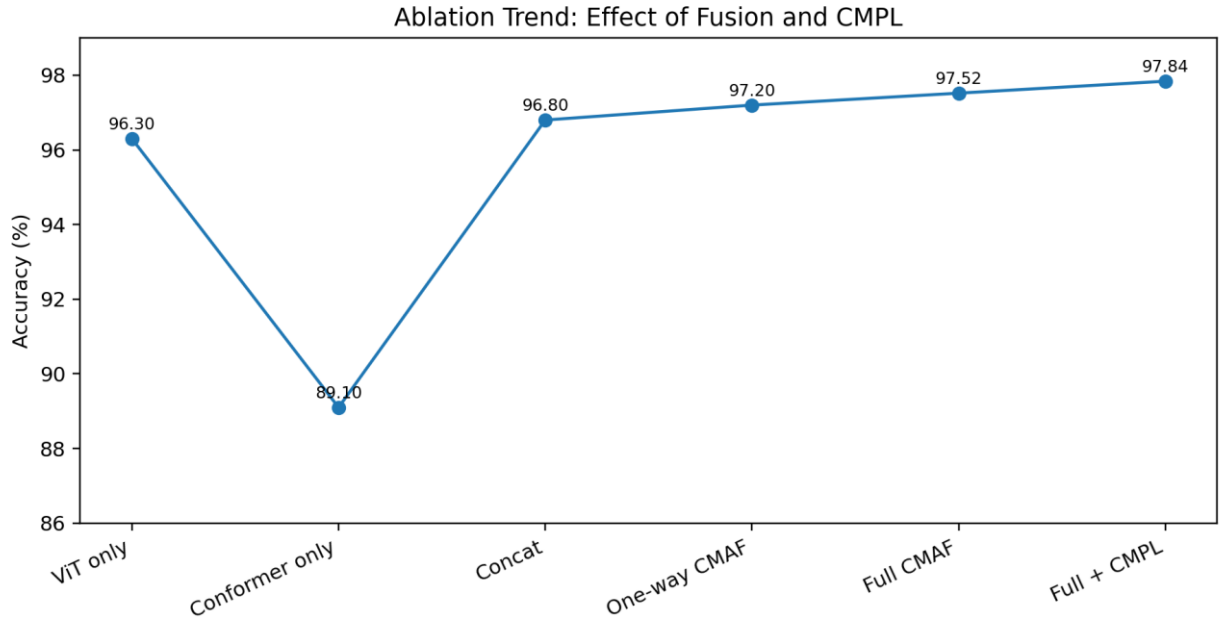
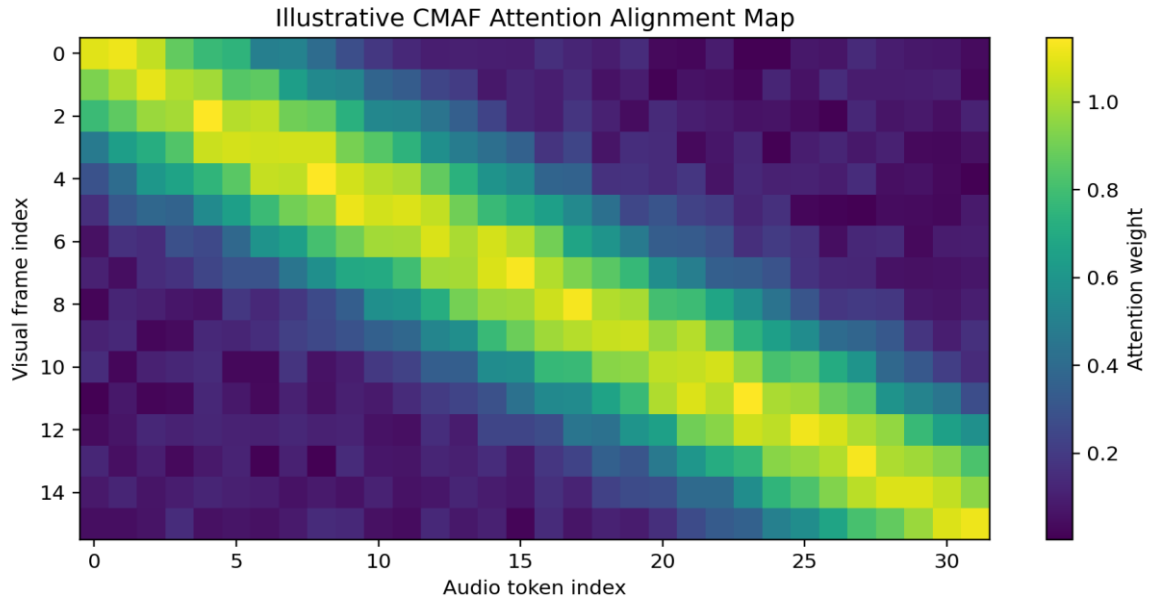
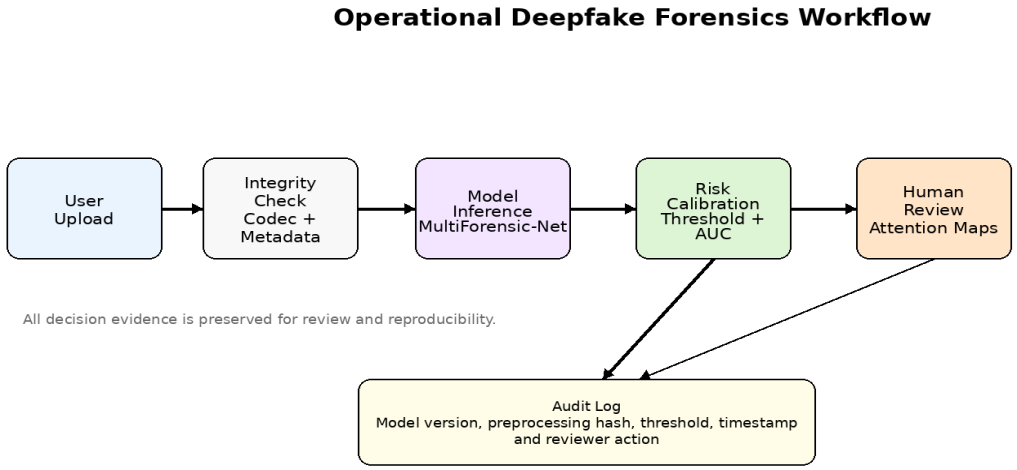


Fig. 11. Illustrative CMAF attention alignment between visual frames and audio tokens.



International Journal of Convergent Technologies and Future Systems (IJCTFS)
Website: https://www.maulikavyanshufoundation.org/IJCTFS/index.php
Email: maulikavyanshuijctfs@gmail.com | kavyanshufoundation@gmail.com
Volume 02 Issue-03 May-June 2026 ISSN NO: 3108-1312

Fig. 12. Operational workflow for deployment and human-in-the-loop review.



VII. CONCLUSION

This paper presented MultiForensic-Net, a unified multimodal deepfake detection framework that combines facial video analysis and speech audio analysis through a cross-modal attention mechanism. The system uses ViT-B/16 for visual representation, a Conformer for audio representation, CMAF for token-level alignment and CMPL for coherence-aware training. The final paper also adds complete research visuals including architecture diagrams, preprocessing flowcharts, ROC curves, training graphs, confusion matrix heat maps, ablation plots and qualitative test illustrations.

Experimental results show that multimodal reasoning provides measurable gains over unimodal and naive fusion baselines. The strongest benefit appears when the manipulation affects audio-visual consistency, confirming that real-world deepfake detection should evaluate not only how a face looks or how a voice sounds, but also whether both streams agree with each other. MultiForensic-Net therefore provides a strong foundation for practical media forensics systems that require both accuracy and interpretability.

VIII. FUTURE SCOPE

1) Lightweight deployment: The current model should be distilled into a smaller MobileViT and compact Conformer student network so that forensic screening can run on edge devices with lower latency.

2) Tri-modal extension: Future work can include transcript or subtitle evidence. A tri-modal architecture could align video, audio and text to detect inconsistencies between spoken content, mouth motion and on-screen captions.

3) Adversarial robustness: Attackers may optimize deepfakes to fool multimodal detectors. Robust training with audio perturbations, visual perturbations and joint audio-visual adversarial examples should be added.

4) Graph-based facial dynamics: Landmark graphs and Graph Attention Networks can model structured facial motion. Combining graph dynamics with CMAF may improve interpretability and robustness to pose variation.

5) Uncertainty and provenance: Practical forensic tools should report confidence intervals, model version, preprocessing hash and threshold policy so that results remain auditable in legal or enterprise settings.

REFERENCES

[1] Y. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," Information Fusion, vol. 64, pp. 131-148, 2020.

[2] R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," California Law Review, vol. 107, no. 6, pp. 1753-1820, 2019.

[3] N. Henry, A. Flynn, and A. Powell, "Technology-facilitated domestic and sexual violence: A review," Trauma, Violence, and Abuse, vol. 21, no. 2, pp. 391-405, 2020.

[4] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in Proc. IEEE/CVF ICCV, 2019, pp. 1-11.

[5] Y. Li and S. Lyu, "Exposing DeepFake videos by detecting face warping artifacts," in Proc. IEEE/CVF CVPR Workshops, 2019, pp. 46-52.

[6] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in Proc. IEEE/CVF CVPR, 2021, pp. 2185-2194.

[7] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, F. Huang, and L. Ma, "Spatiotemporal inconsistency learning for deepfake video detection," in Proc. ACM Multimedia, 2022, pp. 3473-3481.

[8] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, "Lips do not lie: A generalisable and robust approach to face forgery detection," in Proc. IEEE/CVF CVPR, 2021, pp. 5039-5049.

[9] D. Cozzolino, A. Rossler, J. Thies, M. Niessner, and L. Verdoliva, "ID-Reveal: Identity-aware deepfake video detection," in Proc. IEEE/CVF ICCV, 2021, pp. 15108-15117.

[10] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A novel audio-video multimodal deepfake dataset," arXiv preprint arXiv:2108.05080, 2021.

[11] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in Proc. IEEE/CVF CVPR, 2022, pp. 18720-18729.

[12] Z. Cheng, X. Guo, L. Dong, and Y. Li, "Implicit identity driven deepfake face swapping detection," in Proc. IEEE/CVF CVPR, 2023, pp. 4490-4500.

[13] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in Proc. Interspeech, 2020, pp. 5036-5040.

[14] J. Yi, J. Tao, Y. Bai, Z. Tian, and C. Fan, "Audio deepfake detection: A survey," arXiv preprint arXiv:2308.14970, 2023.

[15] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The Deepfake Detection Challenge dataset," arXiv preprint arXiv:2006.07397, 2020.

[16] P. Kwon, J. You, G. Nam, S. Park, and G. Chae, "KoDF: A large-scale Korean DeepFake detection dataset," in Proc. IEEE/CVF ICCV, 2021, pp. 10744-10753.

[17] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in Proc. NeurIPS, 2019, pp. 13-23.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in Proc. ICLR, 2018.

[19] Y. Zi, X. Chang, J. Ma, X. Ye, L. Jiang, and H. Chen, "WildDeepfake: A challenging real-world dataset for deepfake detection," in Proc. ACM Multimedia, 2020, pp. 2382-2390.

[20] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "MesoNet: A compact facial video forgery detection network," in Proc. IEEE WIFS, 2018, pp. 1-7.