Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Transformer-Based Parallel Fusion of Audio and Visual Cues for Robust Deepfake Video Detection

Kushal Sharma[1]
Department of Computer Science and Engineering
Himachal Pradesh Technical University
Hamirpur, Himachal Pradesh, India

Raman Thakur[2]
Department of Computer Science and Engineering
Career Point University Hamirpur

Nitesh Kumar[2]
Department of Computer Science and Engineering
Career Point University Hamirpur

***Abstract* - Deepfake videos threaten the authenticity of digital media. Conventional approaches for detecting deepfakes typically rely on a single type of input, either visual or audio. These unimodal methods often struggle in real-life scenarios due to noise, compression, and tampering. This study presents a transformer-based multimodal deepfake detection framework that jointly exploits audio and video cues through parallel fusion. For feature extraction, the Audio Spectrogram Transformer (AST) is used for audio and the Vision Transformer (ViT) for video. To effectively capture cross-modal relationships, a cross-attention fusion module is employed in parallel with dynamic weighted fusion, enabling the model to learn complementary information while adaptively assigning importance to each modality. The fused representations are concatenated, and the visual, audio, and fused feature representations are each classified using a multi-layer perceptron. A soft-voting ensemble strategy then combines the predictions of the unimodal and multimodal branches. Experimental results on the FakeAVCeleb dataset show that the proposed model, MAVDFormer, achieves an accuracy of 98.83% and an AUC of 0.9992, demonstrating superior performance compared with existing state-of-the-art methods.**

Keywords—Multimodal deepfake detection, Vision Transformer (ViT), Audio Spectrogram Transformer (AST), Cross-attention, Dynamic Weighted Fusion (DWF), Ensemble learning.

1. INTRODUCTION

Deepfake videos are artificially generated or altered media that utilize deep learning algorithms. With rapid progress in AI technologies, deepfake videos are now so realistic that they can be difficult to distinguish from authentic content [1]. The use of software to modify photos and videos has significantly improved [2]. Deepfakes, although used for entertainment purposes, also pose a threat of creating misinformation and compromising security. Deepfake technology has been used to replace the faces of public figures—for example, there have been cases where the face of former Argentine President Mauricio Macri was digitally replaced with that of Adolf Hitler, or where Angela Merkel's appearance was digitally altered to resemble Donald Trump [2]. The origin of such content can be traced back to 2017, when the face of a pornographic actress was replaced with that of a celebrity [3]. These are instances of how easily one's identity can be compromised.

Approaches to detect deepfakes can be either single-modality-based or multimodal-based. Most detection strategies available to date concentrate solely on a single modality, such as an audio signal or a video signal. Visual artifacts of deepfake videos include misaligned features, blurry edges, and inappropriate lighting or texture. Unnatural expressions, errors in lip synchronization, and irregular eye movements are also considered artifacts. Time variations across frames, e.g., flickering or motion disparities, are further important signs of manipulation [4]. Unlike video deepfakes, audio deepfakes primarily involve cloning voices to produce speech that the individual may never have actually said [5]. The evolution of technologies such as text-to-speech (TTS) and voice conversion has facilitated the production of high-quality synthetic human speech [6]. Voice cloning uses a neural network approach that takes an audio sample of someone's voice along with text, and produces human-like voice output from it. A massive number of audio samples are streamed via the Internet every day, but detecting their authenticity remains a difficult task [7].

In spite of the efforts made by researchers, deepfake detection continues to face significant challenges. In particular, detecting human-synthesized voice alongside other facial deepfakes poses a problem. Another challenge relates to the lack of comprehensive manipulation datasets. A reliable detection system would be highly valuable to people who become victims of the misuse of deepfake technology. Detection of deepfakes means classification of images, videos, and audio as either real or fake, and for this, features must be extracted that can distinguish manipulated data from authentic data [8].

Considering these issues, the proposed approach introduces a multimodal deepfake video detection system based on cross-attention modeling and dynamic weighted fusion applied in parallel. Our method uses a Vision Transformer (ViT) network to generate high-level global representations of video frames, while the Audio Spectrogram Transformer (AST) generates discriminative audio features. The cross-attention mechanism is used to expose inconsistent patterns between the visual and audio streams. Simultaneously, dynamic weighted fusion is used to create a complementary multimodal representation, and the resulting features are classified using an MLP-based ensemble.

The main contributions of this work are as follows:

1) A transformer-based multimodal deepfake detection framework (MAVDFormer) integrating a Vision Transformer (ViT) and Audio Spectrogram Transformer (AST) for robust audio-visual feature extraction.

2) A parallel multimodal fusion strategy combining Cross-Attention Fusion (CAF) and Dynamic Weighted Fusion (DWF) to effectively capture complementary cross-modal representations.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

3) A hybrid ensemble classification mechanism that jointly exploits visual, audio, and fused multimodal representations for improved detection performance.

4) Extensive experiments conducted on the FakeAVCeleb dataset demonstrate that the proposed framework outperforms several existing deepfake detection methods.

The remainder of this paper is organized as follows. Section II reviews existing literature on video, audio, and multimodal deepfake detection. Section III presents the proposed MAVDFormer framework, including dataset description, preprocessing, feature extraction, parallel multimodal fusion, classification, and ensemble learning. Section IV describes the experimental setup and discusses the performance evaluation of the proposed framework. Finally, Section V concludes the paper and outlines potential directions for future research.

2. LITERATURE REVIEW

2.1 Video deepfake detection

A large body of work has used visual-only deepfake detection models founded on spatio-temporal inconsistencies in video frames. Studies such as [9], along with several others, utilize CNN-based methods that focus on extracting facial features and pixel-level manipulations to detect inconsistencies, yielding strong results on standard datasets.

[10] proposes a Convolutional Vision Transformer (CViT) that combines CNN and Transformer architectures to detect deepfake videos by learning both local and global features. It emphasizes strong data preprocessing and uses face-based inputs to improve accuracy, achieving around 91.5% accuracy. Similarly, [11] created a dual-attention framework that enhanced facial feature extraction for generated-image detection. [12] created ResViT, and [8] used ViViT to build video deepfake detectors that capture global dependencies and spatio-temporal relations within a frame. [2] surveyed different methods for identifying deepfake videos, focusing on both data-driven and deep-learning-based approaches, and identified dataset constraints, computational complexity, and lack of real-world generalization as major limitations. [13] created a DFT-MF deepfake detector that relies on CNN and mouth/lip-movement analysis to detect speech anomalies; this performed well on the Celeb-DF and Deepfake-TIMIT datasets but was limited by its reliance on mouth-region features. [14] presented an approach to protect a CNN-based deepfake detection framework from adversarial attacks using perturbation techniques, reporting 87.3% accuracy and 76.2% precision, though the framework could not be generalized well because it relied on large datasets. [4] developed a DVDD-LLaMA framework integrating CLIP and Swin Transformers, achieving 67.9% accuracy on the FF++VQA dataset. [15] proposed a CNN-based detector that combines spatial, temporal, and frequency patterns with AdaBoost, achieving 86.5% accuracy on the DFDC dataset, though the model remained susceptible to noisy features. [16] proposed a method using VGG-16 and LSTM to identify deepfake videos from spatial and temporal features, reporting accuracy, precision, and recall of 96.25%, 99.07%, and 93.04% respectively on Celeb-DF. [17] introduced a multi-scale feature fusion (MS-FF) system built on a deep residual network (DRN) for spatial features and an LSTM for temporal modeling, combined with a Diverse Margin Penalty Function for robust face matching, achieving a mean accuracy of 96.5% and a false positive rate of 2.3%.

2.2 Audio deepfake detection

[18] investigated deep learning and CNNs for detecting deepfake audio using the VCTK and LibriSpeech speaker databases. Recordings were preprocessed and converted into images, from which features such as mel-spectrograms and MFCCs were extracted. [19] employed both image-based and feature-based approaches using SVM, STN, and TCN models; the TCN model achieved 92% accuracy on the FoR dataset, though the method's reliance on spectrogram preprocessing limited generalization to raw-audio or multimodal settings. [20] applied MFCC features to train SVM, random forest, and VGG-16 models on the Fake-or-Real dataset, with SVM scoring 98.83% and VGG-16 scoring 93% on challenging subsets, though the approach depended heavily on manual feature extraction. [7] reviewed existing audio deepfake detection models, both machine-learning- and deep-learning-based, and found that most methods convert audio into spectrogram-like features before classification—ML models are accurate but scale poorly, while deep learning models require more complex transformations. [21] proposed a spectrogram-based machine learning approach using ensemble models such as Random Forest, Gradient Boosting, and XGBoost, trained on the Deep-Voice dataset, achieving 99.32% accuracy with XGBoost performing best. [29] introduced AudioFakeNet, a hybrid CNN-LSTM-multi-head-attention architecture that uses MFCC features and attention over informative audio segments, reporting more than 96% accuracy.

2.3 Multimodal deepfake detection

[22] proposed the FakeAVCeleb dataset, comprising videos with real and fake combinations of audio and visual content, and showed through unimodal, ensemble, and multimodal experiments that unimodal methods fail to effectively detect multimodal deepfakes—motivating the need for improved multimodal approaches. [23] proposed a multimodal framework that jointly processes spatial, spectral (DCT-based), and temporal information via LSTM networks, demonstrating that integrating multiple feature types improves detection performance. [9] reviewed deepfake detection methods and highlighted the importance of multimodal fusion combined with CNNs, RNNs, and Transformers. [24] presented a review of 73 sources on deepfake video detection, categorizing techniques into visual, audio, and multimodal approaches and examining datasets such as FaceForensics++, Celeb-DF, and DFDC. MultimodalTrace [5] extracts representations using ResNet-3D for video and ResNet-1D for audio, and fuses them using intra- and inter-modality mixer layers (IAML, IEML), achieving 92.9% accuracy on FakeAVCeleb. [6] proposed a system that processes video with a 3D CNN with attention, audio with a spectrogram CNN, and learns cross-modal relations via cross-attention, fusing all three with a simple ensemble. [25] introduced an architecture that extracts visual cues (blink rate, head pose, facial landmarks) and Mel-spectrogram-based audio features, classified using ANN and VGG19/CNN respectively, flagging a sample as fake if either modality is identified as fake, achieving roughly 94% accuracy. Sultan and Ibrahim (2024) combined VGGFace-with-PCA visual features and MFCC audio features with a LazyClassifier, reporting 95.5% multimodal accuracy and 97.4% audio-only accuracy on a 1,000-video FakeAVCeleb subset. [26] combined visual attention (VGG16-based), audio attention (RNN-based), and cross-modal attention (Bi-GRU-based)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

components, reporting 89% accuracy on FakeAVCeleb. [27] introduced DeepFakeDetNet, fusing texture/shape descriptors (GLCM, XLTP, HOG) from the visual modality with spectral, temporal, and voice-quality audio descriptors, selected via an improved starfish optimization algorithm and classified with a DCNN-LSTM model fused at the score level, achieving 97.80% accuracy, 98.29% recall, 97.10% precision, and 97.70% F1-score on FakeAVCeleb.

3. PROPOSED METHODOLOGY

This chapter presents the proposed MAVDFormer (Transformer-Based Parallel Fusion of Audio and Visual Cues) framework for detecting deepfake videos. The proposed framework is designed to exploit complementary information from both modalities to enhance detection performance. As illustrated in Fig. 1, the framework consists of five major stages: dataset description, data preprocessing, feature extraction, parallel multimodal fusion, classification, and ensemble learning.

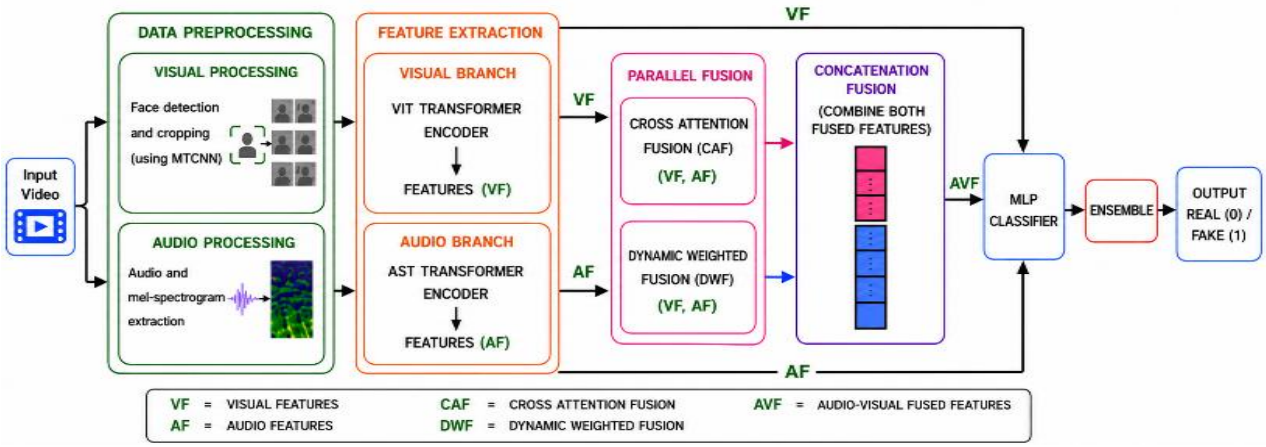


Fig. 1. Proposed MAVDFormer framework for multimodal audio-video deepfake detection.

3.1 Dataset description

FakeAVCeleb [28]: a multimodal deepfake dataset based on audio-visual forgery, with binary labels of authenticity (real vs. fake). It comprises authentic and fake videos in which the face or voice of an individual has been distorted, making it well-suited for studying cross-modal discrepancies. FakeAVCeleb is composed of four categories: Fake Video with Fake Audio (FVFA), Fake Video with Real Audio (FVRA), Real Video with Fake Audio (RVFA), and Real Video with Real Audio (RVRA).

Fig. 2. (a) Sample of a fake video-audio pair from FakeAVCeleb.

Fig. 3. (b) Sample of a real video-audio pair from FakeAVCeleb.

Because FakeAVCeleb contains a limited number of real videos, additional real samples from the LAV-DF dataset were incorporated to improve class balance and generalizability. The dataset was partitioned into training, validation, and testing subsets as shown in Table I, using a 70:30 training-to-testing split, with 10% of the training data reserved for validation. Validation was used to tune parameters, detect overfitting, and apply early stopping, while the held-out test set was used only for final evaluation. As shown in Table II, class distribution is balanced across both the training and test partitions.

TABLE I. DATASET SPLIT FOR MULTIMODAL DEEPFAKE VIDEO DETECTION

Split	Number of Samples
Training	6995
Validation	699
Test	2999
Total	9994

TABLE II. CLASS DISTRIBUTION OF TRAINING AND TEST SETS

Split	Real (0)	Fake (1)	Total
Training	3500	3495	6995
Test	1500	1499	2999
Total	5000	4994	9994

3.2 Preprocessing

The input to the proposed system is a raw video, from which two streams are processed in parallel: one for visual preprocessing and one for audio preprocessing.

3.2.1 Visual preprocessing

From each input video, 20 frames are extracted at equal time intervals. Face detection is performed using MTCNN, a highly accurate multi-task cascaded convolutional neural network

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

used for locating facial regions. The detected faces from each video are stored as separate cropped frames.

3.2.2 Audio preprocessing

The audio track is extracted from each video using FFmpeg, converted to mono, and resampled to 16 kHz for consistency. Log Mel-spectrogram features are then computed using the Librosa library with 128 Mel bands.

3.3 Feature Extraction

3.3.1 Visual network

Cropped facial images are converted into patch embeddings and fed into the encoder of a Vision Transformer (ViT). Its self-attention mechanism captures global relationships across the face, helping detect subtle changes in texture, expression, and geometric irregularities. This produces a set of visual feature embeddings (VF).

3.3.2 Audio network

Mel-spectrograms are embedded and processed by an Audio Spectrogram Transformer (AST) encoder. The learned representation captures abnormalities related to voice characteristics or cloned speech, producing a set of audio feature embeddings (AF).

3.4 Parallel Multimodal Feature Fusion

3.4.1 Cross-attention network

To combine the feature sets from both modalities, a cross-attention fusion network is used, in which the visual features attend to the audio features and vice versa. This allows the model to capture interactions between modalities and to distinguish discrepancies—for instance, between facial movements and speech content. The output of this branch is referred to as the Cross-Attention Fusion (CAF) output.

Cross-attention is formulated using three parameters: the query matrix Q, the key matrix K, and the value matrix V. The attention operation is defined as:

$$CrossAttention(Q, K, V) = softmax(QK^T / \sqrt{dk})\ V \quad (1)$$

where $Q \in \mathbb{R}^{nq \times dk}$ denotes the query representations from one modality, $K \in \mathbb{R}^{nk \times dk}$ the key representations from the other modality, $V \in \mathbb{R}^{nk \times dv}$ the corresponding value representations, and d_k the dimensionality scaling factor.

3.4.2 Dynamic weighted fusion

In parallel with cross-attention, a dynamic weighted fusion (DWF) block refines the merged feature representation by assigning adaptive weights to the audio and visual features according to their relative significance for a given sample. This allows the network to rely more heavily on the more reliable modality when one stream is noisy, incomplete, or corrupted. The output of this branch is referred to as the Dynamic Weighted Fusion (DWF) output.

3.4.3 Concatenation fusion

The outputs of the Cross-Attention Fusion and Dynamic Weighted Fusion branches are combined through concatenation, producing a unified multimodal vector AVF (Audio-Video Fusion) of 1024 dimensions:

$$AVF = Concat(CAF, DWF) \quad (2)$$

3.5 MLP Classifier

The MLP classifier is a fully connected network used to make the final real/fake decision from the extracted features. It learns complex nonlinear relationships through multiple hidden layers with ReLU activation, and outputs a prediction via a sigmoid layer for binary classification. Rather than operating only on the fused features, the proposed framework processes the visual representation VF, the audio representation AF, and the fused representation AVF through three independent MLP branches, each learning nonlinear representations through fully connected ReLU layers; implementation parameters are listed in Table V.

3.6 Ensemble Learning

Ensemble learning combines multiple predictions to achieve more robust classification. It is employed here to minimize modality-specific errors and improve overall detection accuracy. A soft-voting classifier treats the outputs of the three MLP branches as probabilities rather than binary decisions, and the final prediction is obtained by averaging the class probabilities:

$$Pensemble = (PVF + PAF + PAVF) / 3 \quad (3)$$

where PVF, PAF, and PAVF denote the probabilities predicted by the visual, audio, and audio-visual fusion MLP branches, respectively. A standard binary threshold is then applied: if Pensemble ≥ 0.5, the video is classified as fake; otherwise, it is classified as real.

4. RESULTS

4.1 Implementation environment

The experiments were implemented in Jupyter Notebook and executed on an NVIDIA RTX 3050 GPU. Data preprocessing was performed using MTCNN, FFmpeg, and Librosa, while evaluation metrics were computed using scikit-learn. The parameters used for the proposed architecture are listed in Table V.

4.2 Results and discussion

To assess the contribution of each component of the proposed MAVDFormer framework on the FakeAVCeleb dataset, an ablation study was carried out. Different configurations were evaluated step-by-step by applying Cross-Attention Fusion (CAF), Dynamic Weighted Fusion (DWF), and ensemble learning individually and in combination. As shown in Table III, each component contributes to detection quality, with the complete MAVDFormer framework achieving the highest accuracy of 98.83%. Notably, the unimodal ViT branch alone already achieves a strong 99.17% accuracy, and cross-attention fusion alone (96.83%) slightly underperforms it; the full benefit of the multimodal design becomes apparent only once dynamic weighted fusion and ensembling are combined with cross-attention, underscoring that the gain in this framework comes from the combination of fusion strategies and ensembling together, rather than from cross-attention fusion in isolation.

TABLE III. ABLATION STUDY OF THE PROPOSED MAVDFORMER ON THE FAKEAVCELEB DATASET

Configuration	CAF	DWF	Ensbl.	Acc.(%)
ViT	–	–	–	99.17

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Configuration	CAF	DWF	Ensbl.	Acc.(%)
AST	–	–	–	97.13
ViT+AST+CAF	✓	–	–	96.83
ViT+AST+DWF	–	✓	–	97.17

Configuration	CAF	DWF	Ensbl.	Acc.(%)
ViT+AST+CAF+DWF	✓	✓	–	98.00
MAVDFormer (full)	✓	✓	✓	98.83

Table IV compares the proposed multimodal approach against existing deepfake detection methods across different datasets and modalities. The comparison shows that the proposed multimodal approach outperforms the reported unimodal and multimodal baselines, as it leverages complementary information from two data sources.

TABLE IV. PERFORMANCE COMPARISON OF THE PROPOSED WORK WITH EXISTING WORK

Ref.	Method	Dataset Used	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
[27]	DCNN-LSTM	FakeAVCeleb	97.80	97.10	98.29	97.70	–
[8]	ViViT	Celeb-DF v2	87.18	–	–	92.52	–
[29]	CNN + LSTM + Multi-Head Attention	Fake-or-Real	96.00	95.00	92.00	94.00	–
[26]	Optical Flow + VGGNet (video) + MFCC + RNN (audio)	FakeAVCeleb	89.00	–	–	–	–
[6]	3D-ResAttNet + ResNet-CNN + Cross-modal Network + Ensemble	FakeAVCeleb	94.08	92.91	95.43	94.15	–
Proposed	ViT–AST Ensemble (MAVDFormer)	FakeAVCeleb	98.83	99.66	98.00	98.82	99.2

Fig. 2 and Fig. 3 present the confusion matrix and ROC curve of the proposed MAVDFormer model, respectively. The confusion matrix compares actual and predicted labels using True Positives, True Negatives, False Positives, and False Negatives, revealing both overall accuracy and the nature of the model's errors. Of the 1,500 real samples (class 0), 1,495 were correctly classified as real while 5 were misclassified as fake; of the 1,499 fake samples (class 1), 1,469 were correctly identified as fake while 30 were misclassified as real (Fig. 2). The ROC curve illustrates the trade-off between the true positive rate and the false positive rate across classification thresholds; a curve closer to the top-left corner indicates stronger performance. The proposed model achieves an AUC of 0.9992 (Fig. 3), indicating excellent discriminative performance.

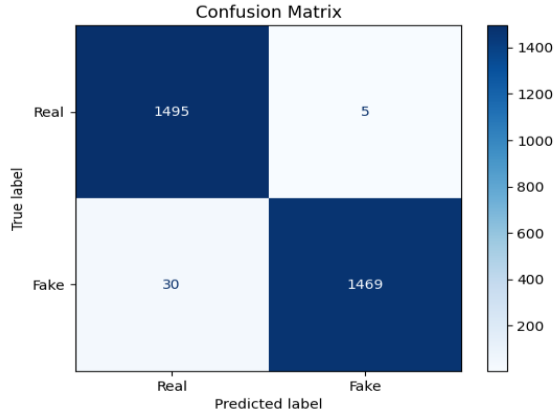
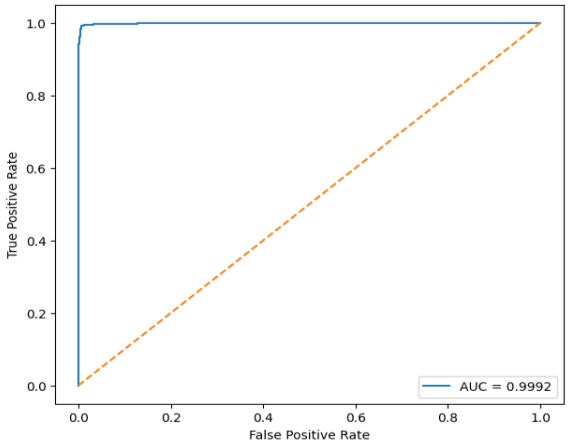


Fig. 4. Confusion matrix of MAVDFormer.



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Fig. 5. ROC curve of MAVDFormer.

TABLE V. PARAMETERS USED FOR IMPLEMENTING THE MODEL

Parameter	Value
Batch size	64
Epochs	50
Early stopping patience	10
Validation split	10%
Optimizer	Adam
Loss function	CrossEntropyLoss
Activation function	ReLU
Hyperparameter tuning (grid search)	{hidden:256, lr:1e-3, dropout:0.3}; {hidden:512, lr:1e-3, dropout:0.5}; {hidden:512, lr:5e-4, dropout:0.4}

5. CONCLUSION AND FUTURE WORK

This work proposed a hybrid multimodal framework, MAVDFormer, that leverages complementary information from the visual and audio domains through parallel fusion. Visual features were extracted using a Vision Transformer (ViT) encoder, while audio features were extracted from Mel-spectrograms using an Audio Spectrogram Transformer (AST). A cross-attention fusion module operating in parallel with dynamic weighted fusion proved effective for combining the two modalities, and an MLP-based ensemble classifier enabled high detection accuracy in distinguishing real videos from deepfakes.

Future work includes training the proposed model on larger and more diverse datasets containing real-world manipulated videos, and exploring alternative fusion and audio-video synchronization strategies, including additional attention mechanisms. A promising direction is the development of real-time deepfake detection suited to social media and live-streaming content. Incorporating explainable AI techniques could further improve the credibility of detection decisions, while additional modalities—such as text transcripts, facial landmark features, physiological signals, and metadata—combined with privacy-preserving and adversarial learning methods, could make the system more reliable, secure, and broadly applicable.

ACKNOWLEDGMENT

The authors declare that no funding was received for conducting this study.

REFERENCES

[1] Y. Patel et al., “Deepfake Generation and Detection: Case Study and Challenges,” IEEE Access, vol. 11, pp. 143296–143323, 2023.

[2] A. Kaur, A. Noori Hoshyar, V. Saikrishna, S. Firmin, and F. Xia, “Deepfake video detection: challenges and opportunities,” Artif. Intell. Rev., vol. 57, no. 6, Jun. 2024.

[3] A. Badale, C. Darekar, L. Castelino, and J. Gomes, “Deep Fake Detection using Neural Networks,” International Journal of Engineering Research & Technology, vol. 9, no. 3, Feb. 2021.

[4] H. Sun, C. Cai, K. A. Lee, L.-P. Chau, and Y. Wang, “Multimodal Large Language Model for Deepfake Video Detection and Description,” in 2025 APSIPA ASC, pp. 2044–2049, Oct. 2025.

[5] M. Anas Raza and K. Mahmood Malik, “Multimodaltrace: Deepfake Detection using Audiovisual Representation Learning,” in 2023 IEEE/CVF CVPRW, Vancouver, BC, Canada, Jun. 2023, pp. 993–1000.

[6] M. Masood, A. Javed, and A. Irtaza, “Attention-based Multimodal learning framework for Generalized Audio-Visual Deepfake Detection,” Oct. 2023.

[7] Z. Almutairi and H. Elgibreen, “A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions,” MDPI Algorithms, May 2022.

[8] K. N. Ramadhani, R. Munir, and N. P. Utama, “Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution and Self Attention,” IEEE Access, vol. 12, pp. 8932–8939, 2024.

[9] G. Gupta, K. Raja, M. Gupta, T. Jan, S. T. Whiteside, and M. Prasad, “A Comprehensive Review of DeepFake Detection Using Advanced Machine Learning and Fusion Methods,” MDPI Electronics, Jan. 2024.

[10] D. Wodajo and S. Atnafu, “Deepfake Video Detection Using Convolutional Vision Transformer,” Mar. 2021, arXiv:2102.11126.

[11] Y. X. Luo and J. L. Chen, “Dual Attention Network Approaches to Face Forgery Video Detection,” IEEE Access, vol. 10, pp. 110754–110760, 2022.

[12] W. Ahmad, I. Ali, S. Adil Shahzad, A. Hashmi, and F. Ghaffar, “ResViT: A Framework for Deepfake Videos Detection,” Int. j. electr. comput. eng. syst., vol. 13, no. 9, pp. 807–813, Dec. 2022.

[13] M. T. Jafar, M. Ababneh, M. Al-Zoube, and A. Elhassan, “Digital Forensics and Analysis of Deepfake Videos,” in 2020 11th ICICS, Apr. 2020, pp. 53–58.

[14] S. Dhesi, L. Fontes, P. Machado, I. K. Ihianle, F. F. Tash, and D. A. Adama, “Mitigating Adversarial Attacks in Deepfake Detection: An Exploration of Perturbation and AI Techniques,” Sep. 2023, arXiv:2302.11704.

[15] B. Thirumaleshwari Devi and R. Rajasekaran, “Deepfake Video Detection Using Ada-Boosting on the DFDC Dataset,” Procedia Computer Science, Elsevier, 2025, pp. 1091–1101.

[16] L. Boongasame, J. Boonpluk, S. Soponmanee, J. Muangprathub, and K. Thammarak, “Design and Implement of Deepfake Video Detection Using VGG-16 and Long Short-Term Memory,” Applied Computational Intelligence and Soft Computing, vol. 2024, 2024.

[17] G. Yogarajan, S. Soundhariya, and R. S. Harini, “Robust deepfake detection using multi-scale feature fusion,” Multimed. Tools Appl., vol. 84, no. 33, pp. 40959–40981, Oct. 2025.

[18] M. McUba, A. Singh, R. A. Ikuesan, and H. Venter, “The effect of deep learning methods on deepfake audio detection for digital investigation,” Procedia Computer Science, Elsevier, 2023, pp. 211–219.

[19] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar, and F. Kazi, “A Deep Learning Framework for Audio Deepfake Detection,” Arab. J. Sci. Eng., vol. 47, no. 3, pp. 3447–3458, Mar. 2022.

[20] A. Hamza et al., “Deepfake Audio Detection via MFCC Features Using Machine Learning,” IEEE Access, vol. 10, pp. 134018–134028, 2022.

[21] R. Bohara and A. K. Bairwa, “Detecting Deepfake Audio Using Spectrogram-Based Machine Learning Approaches,” IEEE Access, vol. 13, pp. 149478–149489, 2025.

[22] H. Khalid, M. Kim, S. Tariq, and S. S. Woo, “Evaluation of an Audio-Video Multimodal Deepfake Dataset using Unimodal and Multimodal Detectors,” in ADGD 2021, ACM MM 2021, pp. 7–15.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

[23] J. K. Lewis et al., "Deepfake video detection based on spatial, spectral, and temporal inconsistencies using multimodal deep learning," in Proc. AIPR, Oct. 2020.

[24] M. Alrashoud, "Deepfake video detection methods, approaches, and challenges," Elsevier, Jun. 2025.

[25] K. Gandhi, P. Kulkarni, T. Shah, P. Chaudhari, M. Narvekar, and K. Ghag, "A Multimodal Framework for Deepfake Detection," Oct. 2024.

[26] S. Asha, P. Vinod, and V. G. Menon, "A defensive attention mechanism to detect deepfake content across multiple modalities," Multimed. Syst., vol. 30, no. 1, Feb. 2024.

[27] S. Bahadure and V. Mane, "Deepfake detection using deep convolutional neural network and long short-term memory," Multimed. Tools Appl., vol. 85, no. 3, p. 259, Mar. 2026.

[28] H. Khalid, S. Tariq, M. Kim, and S. S. Woo, "FakeAVCeleb: A Novel Audio-Video Multimodal Deepfake Dataset," arXiv:2108.05080, 2021.

[29] S. Dilbar, M. A. Qureshi, S. K. Noon, and A. Mannan, "AudioFakeNet: A Model for Reliable Speaker Verification in Deepfake Audio," Algorithms, vol. 18, no. 11, Nov. 2025.