Journal of Applied Artificial Intelligence
https://jaai.sabapub.com
ISSN: 2709-5908

2026 Volume 7, Issue 1: 87–102

DOI: https://doi.org/10.48185/jaai.v7i1.1939

End-to-End Multimodal Emotion Recognition with Deep Temporal and Cross-Modal Feature Integration

Rexcharles Enyinna Donatus [1, 5*], Oludele Awodele[2], Osondu E. Oguike[3] and Amina Sambo-Magaji [4]

[1]Africa Centre of Excellence on Technology Enhanced Learning, National Open University of Nigeria, Abuja and 900108, Nigeria
[2]Department of Computer Science, Babcock University, Ilishan-Remo and 121103, Ogun, Nigeria.
[3]Department of Computer Science, University of Nigeria, Nsukka and 410101, Enugu, Nigeria.
[4] Digital Literacy & Capacity Development Department, National Information Technology Development Agency, Abuja and 900104, Nigeria.
[5]Aerospace Engineering Department, Air Force Institute of Technology, Kaduna and 800283, Nigeria.

Received: 21.12.2025 • Accepted: 28.03.2026 • Published: 04.06.2026 • Final Version: 30.06.2026

Abstract: Reliable emotion recognition under real-world conditions remains challenging due to noise, occlusion, and subtle or overlapping affective cues. To address these limitations, this study presents an end-to-end temporal multimodal framework that jointly models facial and vocal expressions. The approach integrates deep residual networks for spatial and spectral feature extraction with bidirectional LSTM networks for sequence-level temporal modeling, while attention mechanisms refine modality-specific features and align audio–visual representations prior to classification. Audio signals are represented using Mel-Frequency Cepstral Coefficients (MFCC), and facial features are extracted from video frames, both processed through a shared ResNet-50 backbone. The framework was evaluated on the RAVDESS and CREMA-D datasets under strict subject-disjoint cross-validation, achieving classification accuracies of 91.22% and 87.32%, respectively. Confusion-matrix analyses indicate improved discrimination across emotionally overlapping categories, demonstrating that structured integration of spatial encoding, temporal modeling, and cross-modal attention yields robust emotion recognition. These results highlight the framework's potential for real-world affective computing applications and provide a foundation for future research in multimodal emotion-aware systems.

Keywords: Audio–Visual Fusion, Bidirectional LSTM, Cross-Modal Feature Integration, Emotion Recognition, Temporal Modeling.

1. Introduction

Emotion recognition is a core capability for human-centered artificial intelligence, enabling computational systems to perceive, interpret, and respond to affective states in ways that are socially appropriate and contextually meaningful [1-3]. Accurate emotion understanding underpins a wide range of applications, including natural human–computer interaction, adaptive educational technologies that respond to learner engagement or frustration, affect-aware healthcare monitoring, and intelligent surveillance and driver assistance systems [4-7]. As digital platforms increasingly

* Corresponding Author: charlly4eyims@yahoo.com

mediate communication, learning, and healthcare delivery, robust emotion inference from subtle and dynamically evolving cues has become critical for trust, safety, and long-term user acceptance [8-9].

Despite substantial progress, reliable emotion recognition in real-world conditions remains challenging. Emotional cues are often noisy, incomplete, or ambiguous due to background interference, occlusion, sensor limitations, and individual or cultural variability. In many practical scenarios, facial expressions may be partially obscured or deliberately controlled, while vocal prosody can be degraded by environmental noise or recording conditions [5, 8, 10]. Moreover, emotional states frequently overlap in their observable manifestations, making it difficult to distinguish between closely related affective categories such as confusion, frustration, or mild anger. These factors expose the limitations of models trained and evaluated under constrained laboratory settings, which often fail to generalize to unconstrained human–machine interaction environments [4, 9].

Early emotion recognition systems relied on handcrafted features combined with conventional classifiers such as support vector machines. While effective in controlled environments, these approaches lacked robustness and scalability [5, 11]. The adoption of deep learning has significantly advanced the field by enabling automatic extraction of discriminative representations from speech and facial data. Convolutional neural networks and transfer learning from large-scale pretrained models, particularly residual networks, have proven highly effective for facial expression recognition, while recurrent and convolutional architectures applied to spectral features such as mel-frequency cepstral coefficients (MFCCs) have achieved strong performance in speech emotion recognition [12-15].. Nevertheless, unimodal systems remain inherently fragile when their single modality is degraded or when affective cues are subtle or conflicting [5, 8, 16].

To address these limitations, multimodal emotion recognition has emerged as a dominant paradigm, integrating complementary information from audio, visual, and sometimes textual or physiological signals. Numerous studies have demonstrated that combining modalities yields improved accuracy and robustness compared to unimodal systems, particularly in the presence of incomplete or ambiguous cues [4, 17-18]. Audio and visual modalities are especially complementary: facial dynamics convey fine-grained valence and expressive detail, while speech prosody and temporal rhythm provide cues related to arousal and emotional intensity [6, 8]. However, a substantial proportion of existing multimodal systems rely on static fusion strategies, such as simple feature concatenation or late decision-level fusion, which do not explicitly model time-varying interactions between modalities [4, 17].

Emotion is inherently temporal, unfolding through dynamic facial movements, speech articulation, and evolving conversational context. Models that operate on isolated frames or short segments are therefore limited in their ability to capture long-range dependencies, such as sustained affective states or gradual emotional transitions. This has motivated the use of sequence modeling architectures, particularly long short-term memory and bidirectional LSTM networks, which have shown strong performance in speech emotion recognition and, to a lesser extent, in video-based affect analysis [13-14]. More recently, attention-based and Transformer architectures have been introduced to selectively emphasize emotionally salient time steps and features [10, 18]. Nevertheless, existing systems often emphasize either temporal modeling or multimodal fusion, but rarely integrate both within a unified and balanced framework**.**

Early deep learning approaches to audio–visual emotion recognition typically relied on independent modality-specific encoders followed by high-level fusion. [19] introduced one of the first hybrid deep models combining a 2D CNN for audio and a 3D CNN for visual streams, with feature-level fusion performed using a deep belief network. While this approach demonstrated the

feasibility of deep multimodal learning, temporal dynamics were modeled implicitly through segment pooling, and cross-modal relationships were not explicitly structured. Similarly, [20] processed Mel spectrograms and selected facial frames using separate CNNs and applied late fusion via stacked extreme learning machines and SVMs. Although effective on acted datasets, this decision-level fusion strategy limited fine-grained temporal alignment and inter-modal reasoning.

Subsequent works sought to improve temporal modeling by integrating recurrent architectures. [21] combined a 3D CNN for visual features with a CNN–RNN pipeline for audio spectrograms, followed by high-level fusion using a biologically inspired emotional learning model. Temporal dependencies were captured independently within each modality, but fusion remained shallow, relying on a joint classifier rather than explicit cross-modal interaction. In parallel, unimodal speech emotion recognition studies established MFCC as a robust acoustic representation [22].The authors in [23] demonstrated that deep CNNs trained on MFCC-based spectrograms achieved strong performance, but temporal relationships were embedded implicitly within fixed-size representations rather than modeled as sequences.

More recent research emphasized stronger backbones and improved fusion design. [24] employed deep CNN encoders for facial expressions and VGGish embeddings for audio, followed by RNN-based temporal modeling on concatenated features. Although temporal dependencies were better handled, fusion was still performed through concatenation, without mechanisms to selectively align or weight modality-specific cues. [17] similarly adopted model-level fusion of independently trained audio and visual networks and reported high accuracies on RAVDESS. However, fusion was static, and evaluation protocols were not consistently subject-disjoint, limiting conclusions about generalization.

A major shift occurred with the introduction of attention-based and cross-modal fusion mechanisms. [25] proposed a fully attention-driven architecture incorporating spatial, channel, and temporal attention within CNN-based encoders and an explicit audio–visual cross-attention fusion layer. This work demonstrated the effectiveness of modeling inter-modal dependencies under subject-independent evaluation on RAVDESS and CREMA-D. However, temporal dynamics were primarily captured through short-range convolutional receptive fields and local attention windows, constraining the modeling of long-range emotional trajectories

In parallel, unimodal and hybrid audio models continued to advance temporal reasoning. [26] showed that combining CNNs on Mel spectrograms with BiLSTM networks on MFCC features, augmented by multiple attention mechanisms, substantially improved speech emotion recognition. Their results highlight the effectiveness of BiLSTM-based temporal modeling and attention for capturing long-range dependencies, although the framework remained audio-only and did not address multimodal interaction.

Very recent studies have explored Transformer-based fusion strategies [1]. [27] introduced coordination attention transformers that dynamically balance self- and cross-modal attention, while [28] proposed an adaptive Transformer-based cross-modal fusion network. These methods demonstrate strong performance gains, but they rely heavily on Transformer stacks and do not exploit the inductive bias of recurrent models such as BiLSTM for sequential emotional dynamics. Moreover, visual encoders in these works are often generic CNNs rather than explicitly leveraging deep residual representations optimized for facial affect**.**

In summary, prior studies have demonstrated the effectiveness of deep convolutional backbones, MFCC-based acoustic representations, recurrent temporal modeling, and attention-driven fusion for emotion recognition. However, these components are most often investigated independently or combined in loosely coupled pipelines, resulting in inconsistent temporal representations across

modalities and limited cross-modal interaction. There remains a gap for end-to-end frameworks that jointly and coherently integrate deep visual encoding, MFCC-centered audio modeling, sequence-level temporal learning, and adaptive multimodal fusion under rigorous subject-independent evaluation.

This study addresses this gap by proposing a unified temporal multimodal emotion recognition framework that systematically integrates deep spatial encoding, sequence-level temporal modeling, and feature-level cross-modal fusion within a single architecture. Pretrained residual networks are employed to extract high-level spatial representations from both audio-derived and visual inputs, while bidirectional LSTM networks are applied consistently to model temporal dynamics in each modality. Crucially, cross-modal attention is introduced after temporal encoding, enabling adaptive alignment of audio and visual feature sequences prior to classification.

While the individual components of the framework such as ResNet backbones, BiLSTM temporal modeling, and attention mechanisms are well established, the novelty of this work lies in their structured organization and joint optimization. Unlike many existing approaches that apply temporal modeling to only one modality or rely on static feature concatenation, the proposed framework enforces a shared sequence-level temporal abstraction for both audio and visual streams before fusion. This design ensures that cross-modal attention operates on temporally aligned and semantically comparable representations, allowing the model to capture complementary affective cues more effectively.

Furthermore, fusion is treated not as a late-stage feature combination step but as an explicit alignment process driven by cross-modal attention. This enables the framework to dynamically emphasize informative modality-specific cues over time, improving discrimination in emotionally ambiguous and overlapping categories. The effectiveness of this design is validated through subject-disjoint cross-validation on two benchmark datasets, RAVDESS and CREMA-D, providing a robust assessment of generalization across speakers and expressive styles.

The main contributions of this work can be summarized as follows:

1. An end-to-end temporal multimodal architecture that coherently integrates deep spatial encoding, sequence-level temporal modeling, and attention-guided feature fusion.
2. A unified temporal learning strategy applied consistently to both audio and visual modalities, enabling aligned and comparable temporal representations.
3. A feature-level cross-modal attention mechanism that explicitly aligns temporally encoded audio and visual features prior to classification.
4. A comprehensive experimental evaluation under strict subject-disjoint protocols, including detailed class-wise and confusion-matrix-based analyses on RAVDESS and CREMA-D.

Together, these contributions demonstrate that improved multimodal emotion recognition performance arises not from introducing new individual components, but from their principled integration into a coherent temporal and cross-modal learning framework.

2. Methodology

This section describes the methodological framework adopted for the proposed temporal multimodal emotion recognition system. It details the data preparation procedures, model architecture, and fusion strategy employed to integrate audio and visual information.

2.1. Dataset

In this study, experiments are conducted on two widely benchmark datasets, RAVDESS and CREMA-D, using strictly audio-visual recordings to ensure multimodal consistency.

RAVDESS is a gender-balanced multimodal dataset comprising 1,440 audio-visual clips from 24 professional actors, covering eight discrete emotions with controlled intensity variations and an average duration of approximately 3.82 ± 0.34 seconds. CREMA-D consists of 7,440 audio-visual clips recorded by 91 actors of diverse ages and ethnic backgrounds, expressing six basic emotions across multiple intensity levels, with an average duration of about 3.63 ± 0.53 seconds [25]. Only the audio-visual modality is used in both datasets to align with the multimodal objective of this work.

Despite their widespread adoption, prior studies using these datasets often rely on inconsistent or actor-dependent evaluation protocols, which can lead to inflated performance estimates. To address this limitation, the present study adopts a subject-disjoint evaluation strategy, enabling a more reliable assessment of generalization across unseen speakers and ensuring fair comparison with recent state-of-the-art approaches. Figure 1 and Figure 2 present representative facial video frames extracted from the RAVDESS and CREMA-D datasets, respectively, illustrating typical emotional expressions captured across different speakers and recording conditions.

Figure 1. Representative video frames of different facial emotions from the RAVDESS dataset.

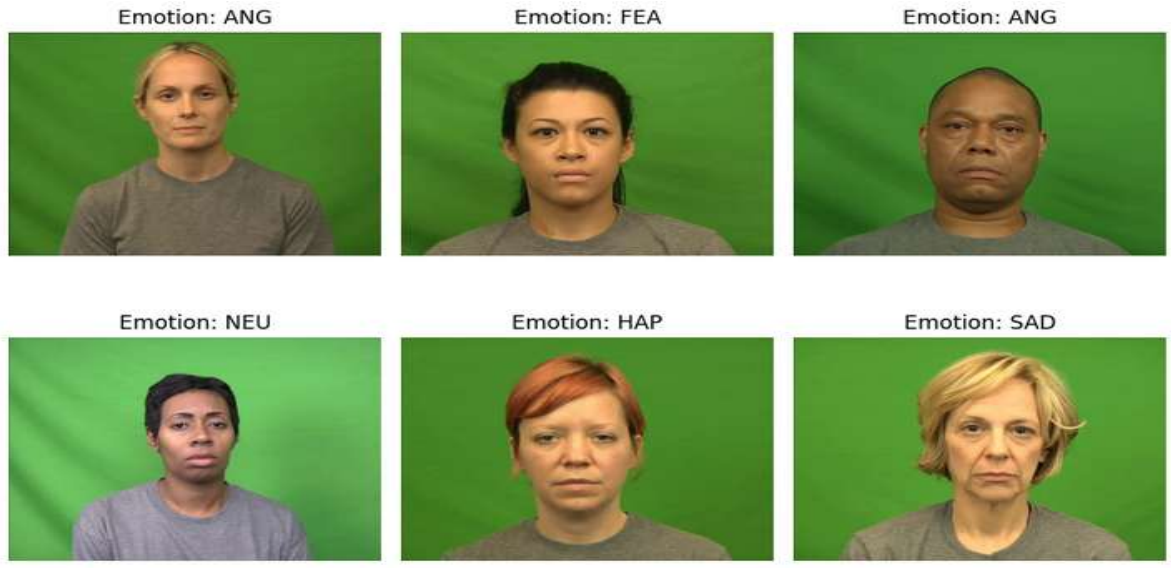

Figure 2. Representative video frames of different facial emotions from the CREMA-D dataset.

2.2. Implementation Details

Audio processing was carried out using Librosa, while OpenCV was employed for video frame extraction and face preprocessing. All experiments for the proposed multimodal emotion recognition framework were conducted on a workstation equipped with an NVIDIA RTX 3090 GPU with 24 GB

of dedicated memory and an Intel Core i7-class processor, running on a Windows operating system. Model development and training were implemented using the TensorFlow deep learning framework with the Keras API, enabling efficient GPU utilization and reproducible experimentation. All experiments were executed using Python version 3.10, with fixed random seeds applied to ensure consistent and repeatable results. Audio and visual streams were processed independently prior to multimodal fusion, following a parallel design that preserves modality-specific temporal structure.

2.3. Framework Overview

Inspired by the architectural principles outlined in [2, 15], the proposed framework is organized around two fundamental design components: a multimodal feature fusion mechanism for integrating complementary audio–visual cues, and a recurrent temporal modeling module based on Bidirectional Long Short-Term Memory (BiLSTM) networks to capture the dynamic evolution of emotional expressions over time. The model operates on high-level embeddings obtained from the preprocessing and feature extraction procedures described in Section 2.4, ensuring consistent and comparable representations across modalities prior to temporal encoding and fusion.

The overall architecture follows a parallel-stream design as shown in fig. 3, where audio and visual modalities are processed independently through dedicated feature extraction and temporal modeling pipelines. Section 2.4 details the data extraction and preprocessing steps for both modalities, while Section 2.5 presents the architecture of the proposed model, including the attention-guided multimodal fusion strategy employed in this study.

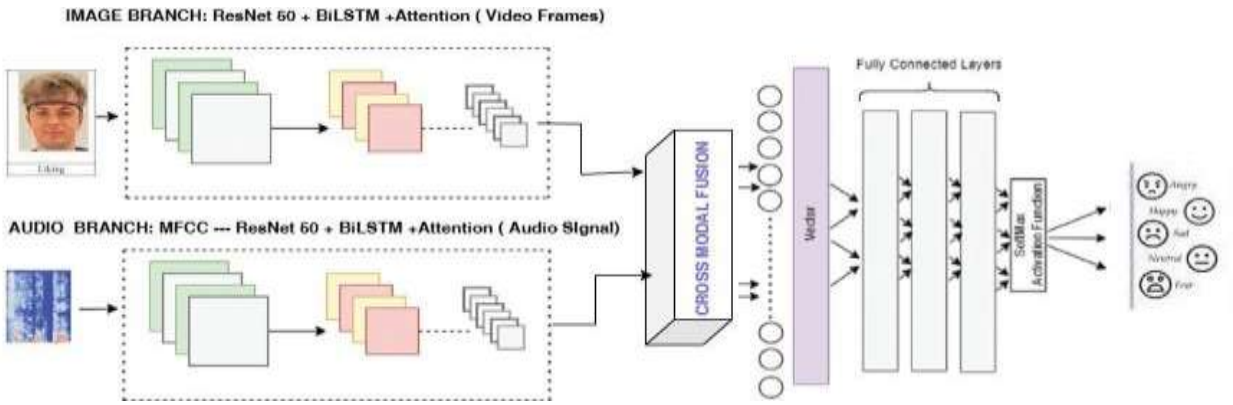

Figure 3. Overview of the Proposed Attention-Based Multimodal Framework for Temporal Emotion Recognition

2.4. Feature Extraction

Audio Feature Extraction

For the audio modality, speech signals were first extracted directly from the audio–visual recordings and resampled to 16 kHz to ensure consistency across samples and compatibility with standard speech processing pipelines. Short-time analysis was then performed using a sliding window of 25 ms with a hop size of 10 ms, allowing fine-grained temporal resolution while preserving relevant prosodic dynamics.

From each frame, Mel-Frequency Cepstral Coefficients (MFCCs) were computed to capture the perceptually relevant spectral characteristics of speech [15]. In addition to the static MFCCs, first- and second-order temporal derivatives were extracted to model dynamic changes in vocal articulation and prosody. These coefficients were stacked to form a three-channel time–frequency representation, encoding both spectral structure and temporal variation within the speech signal.

Rather than employing a task-specific shallow encoder, the MFCC representations were treated as image-like inputs and processed using a ResNet-50 backbone pretrained on ImageNet and fine-tuned for emotion recognition, enabling robust learning of local spectral patterns under speaker variability and noise. The residual architecture further facilitates stable gradient propagation when learning discriminative emotional patterns from high-dimensional acoustic representations.

The output of the global average pooling layer of the ResNet-50 network yields a 2048-dimensional embedding for each temporal segment, providing a compact yet expressive high-level representation of the audio modality. These embeddings are temporally aligned with the corresponding visual features and subsequently passed to the Bidirectional LSTM network for sequence-level modeling of emotional dynamics.

Figure 4 illustrates example MFCC representations extracted from RAVDESS audio samples, highlighting distinct spectral patterns associated with different emotional categories, including happiness, sadness, anger, neutrality, fear, and disgust.

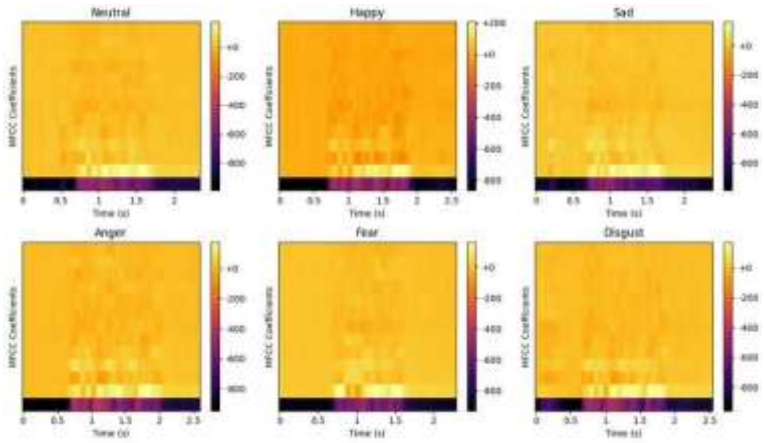

Figure 4. MFCC Spectrogram Extracted from RAVDESS Audio Sample for Emotion Classification.

Visual Feature Extraction

For the visual modality, facial information was extracted from video sequences using a frame-based processing pipeline implemented with OpenCV. Each video was first decomposed into individual frames at a fixed sampling rate to preserve temporal consistency across samples. Face localization was performed on each frame using OpenCV's built-in face detection utilities, which are well suited for controlled datasets such as RAVDESS and CREMA-D, where subjects are recorded in frontal poses under stable illumination conditions. Given the consistent framing and limited background clutter in both datasets, this approach provides reliable face detection without the computational overhead associated with more complex deep-learning-based detectors.

Following face detection, the facial region was cropped from each frame and resized to a fixed spatial resolution of 224 × 224 pixels to match the input requirements of the subsequent deep neural network. Pixel intensities were normalized to ensure numerical stability during training [29]. No explicit landmark alignment was applied, as the datasets exhibit minimal pose variation and consistent facial orientation across samples. This design choice allows the model to learn discriminative facial representations directly from appearance cues rather than relying on handcrafted alignment procedures.

Each preprocessed facial frame was then passed through a ResNet-50 backbone pretrained on ImageNet to extract high-level spatial features. The output of the global average pooling layer was used as a compact 2048-dimensional representation for each frame. These frame-level embeddings were temporally ordered to form a visual feature sequence corresponding to each video clip, which was subsequently used for temporal modeling via bidirectional LSTM networks. This pipeline ensures that both spatial facial characteristics and their temporal evolution are effectively captured while maintaining computational efficiency and architectural consistency with the audio processing stream.

2.5 Temporal Modeling and Multimodal Fusion

Temporal Modeling

Temporal dependencies in both modalities were modeled independently using Bidirectional Long Short-Term Memory (BiLSTM) networks. Each BiLSTM layer consists of 128 hidden units per direction, resulting in a 256-dimensional temporal representation per modality. This configuration provides strong bidirectional context while maintaining training stability. Layer Normalization was applied to the BiLSTM outputs to improve convergence and ensure robustness for variable-length sequences.

Fusion and Classification

Following modality-specific temporal encoding, the proposed framework performs attention-guided feature-level fusion to integrate audio and visual representations in a principled and adaptive manner. Rather than relying on static fusion strategies, the model explicitly structures both intra-modal refinement and inter-modal alignment prior to classification.

First, temporal self-attention mechanisms are applied independently to the temporally encoded audio and visual feature sequences. This operation reweights time steps within each modality based on their relative contribution to emotional expression, allowing the model to emphasize salient temporal segments while suppressing redundant or weakly informative intervals. By operating after BiLSTM-based temporal modeling, self-attention refines sequence representations without disrupting learned temporal dependencies.

The refined modality-specific representations are subsequently passed to a cross-modal attention module, which serves as the core fusion mechanism. This module explicitly models interactions between audio and visual streams by aligning temporally encoded features and dynamically weighting their contributions during fusion [30]. Unlike simple concatenation, which assumes equal reliability across modalities and time, cross-modal attention enables the framework to selectively prioritize complementary cues while attenuating modality-specific noise. This design is particularly important in emotion recognition, where expressive intensity and timing often differ between facial movements and vocal prosody.

The resulting fused representation captures coherent audio–visual emotional dynamics and is subsequently passed through fully connected layers with ReLU activation. A final Softmax layer produces categorical emotion predictions. The network was optimized using the Adam optimizer with a fixed learning rate of 1×10^{-4} and categorical cross-entropy loss. Training is conducted for 200 epochs with a batch size of 32. Dropout with a rate of 0.4 is applied to both the BiLSTM outputs and dense layers to mitigate overfitting and improve generalization.

Importantly, no data augmentation is employed in this study. This design choice ensures that observed performance gains arise from the proposed fusion architecture and attention mechanisms rather than artificial sample inflation. Overall, the fusion and classification strategy emphasizes

structured cross-modal interaction, enabling robust emotion recognition under subject-disjoint evaluation conditions.

Table 1 summarizes the key hyperparameters used for training. These values were selected based on empirical validation and established practices in multimodal emotion recognition to ensure stable convergence and reliable accuracy-based evaluation under subject-disjoint protocols.

Table 1. Selected Hyperparameters for the Proposed Framework

Hyperparameter	Fixed Value	Rationale
Optimizer	Adam (lr = 1e-4)	Stable convergence for deep multimodal models
Loss Function	Categorical Cross-Entropy	Standard for multi-class emotion recognition
Batch Size	32	Balance between efficiency and stability
Epochs	200	Ensures convergence under early stopping
BiLSTM Units	128 per direction	Strong temporal modeling without overfitting
Dropout Rate	0.4	Regularization for BiLSTM and FC layers
Attention Type	BiLSTM-based temporal and cross-modal attention	Adaptive weighting of temporal and cross-modal features
Fusion Strategy	Feature-level with Attention	Preserves complementary audio–visual cues

This implementation ensures a transparent, reproducible pipeline that integrates deep spatial encoding, temporal sequence modeling, and principled multimodal fusion for robust emotion recognition.

2.6. Evaluation Protocol

Model performance was evaluated using overall classification accuracy and confusion-matrix-driven class-wise analysis under strict subject-disjoint cross-validation. This evaluation strategy provides a transparent assessment of both global recognition performance and class-level behavior, which is particularly important for emotion recognition tasks involving overlapping affective categories. All experiments were conducted under subject-disjoint protocols, ensuring that speakers appearing in the test set were never observed during training and enabling a reliable assessment of generalization across unseen individuals.

3. RESULTS AND DISCUSSION

This section presents a comprehensive evaluation of the proposed audio–visual emotion recognition framework and provides a critical analysis of its performance across datasets, modalities, and architectural components.

3.1. Overall Performance Evaluation

Table 3.1 reports the overall classification accuracy achieved by the proposed framework in comparison with recent multimodal emotion recognition approaches. The proposed method attains 91.22 per cent accuracy on RAVDESS and 87.32 per cent accuracy on CREMA-D, outperforming existing state-of-the-art models evaluated under comparable conditions.

Table 2. Accuracy Comparison Between the Proposed Method and Existing Approaches

Method / Study	**YEAR**	**RAVDESS Accuracy (%)**	**CREMA-D Accuracy (%)**

Ghaleb *et al.*	2020	79.0	74.0
Mocanu *et al.*	2023	89.25	84.57
Middya *et al.*	2022	86.0	–
Ghaleb *et al.*	2023	79.0	74.0
Developed Method	**2025**	**91.22**	**87.32**

The improvement is particularly notable on CREMA-D, which is widely regarded as more challenging due to its larger speaker pool, broader age distribution, ethnic diversity, and variation in emotional intensity. These factors introduce significant intra-class variability, making high performance under subject-disjoint evaluation especially difficult. The observed gains therefore indicate that the proposed framework generalizes effectively beyond speaker-specific cues.

The consistent superiority of the multimodal system over unimodal baselines reinforces a key observation in affective computing, emotion is rarely expressed reliably through a single modality**.** Facial expressions and vocal prosody provide complementary affective information, allowing one stream to compensate when the other is ambiguous, suppressed, or noisy.

Unlike static fusion strategies, the proposed framework performs feature-level fusion after temporal modeling, enabling effective alignment of asynchronous audio and visual cues. This design choice proves critical for emotions such as fear and sadness, where facial and vocal expressions often peak at different times.

3.2. Confusion Matrix and Class-Wise Analysis

Figure 5 and Figure 6 present the confusion matrices obtained on RAVDESS and CREMA-D, respectively. Both matrices exhibit strong diagonal dominance, indicating reliable discrimination across most emotion categories.

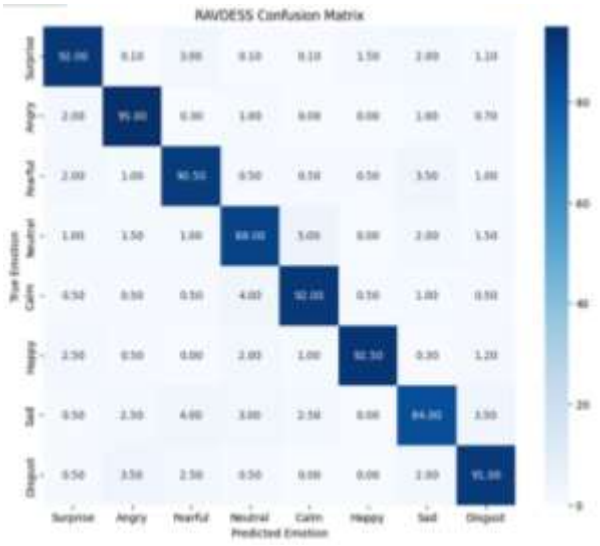

Figure 5. RAVDESS

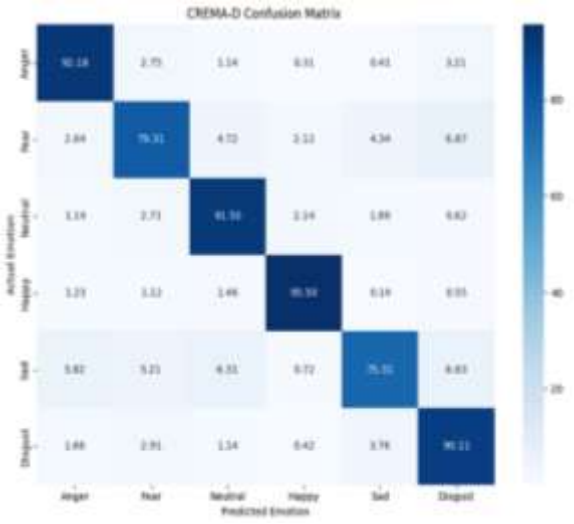

Figure 6. CREMA- D

On RAVDESS**,** the highest recognition rates are observed for happiness and anger**,** each exceeding 92 percent true positive rate. These emotions are characterized by pronounced facial movements and distinct vocal patterns, making them easier to detect in both modalities. Emotions such as calm and surprise are also recognized with high consistency, reflecting effective temporal modeling of subtle expression transitions.

Misclassifications on RAVDESS occur primarily between neutral and calm**,** which is expected given their close affective proximity and limited expressive contrast. Importantly, confusion between high-arousal emotions such as anger and fear is minimal, suggesting that the model successfully captures arousal-related temporal cues.

On CREMA-D**,** recognition performance is highest for happiness (95.50 percent) and anger (92.18 percent)**,** consistent with their strong expressive markers. However, increased confusion is observed between fear and sadness, as well as between sadness and neutral**,** reflecting the lower intensity and more nuanced emotional delivery in this dataset. These confusions mirror known human annotation challenges on CREMA-D, where even human raters achieve only 63.6 percent accuracy in the audio–visual setting.

Despite these challenges, the proposed model maintains robust performance across all classes, demonstrating resilience to speaker variability and expressive ambiguity. No systematic performance differences are observed across gender, indicating that the framework does not overfit to gender-specific affective patterns.

Table 3 further confirms these trends by reporting class-wise true positive rates, highlighting consistent performance across emotion categories and reinforcing the reliability of the proposed fusion strategy.

Table 3. shows Class-wise True Positive Rates (TP %) for the Proposed Multimodal Framework.

DATASET	Surprise	Anger	Fear	Neutral	Calm	Happy	Sad	Disgust
RAVDESS	92.00	95.00	90.50	88.00	92.00	92.50	84.00	91.00
CREMA-D	—	92.18	79.31	91.50	—	95.50	75.31	90.11

This outcome aligns with the controlled recording conditions of both datasets, where facial expressions are clearly visible and consistently framed. Facial cues provide strong spatial indicators for emotions such as happiness and surprise, which are often expressed through distinctive muscle movements.

Nevertheless, These trends are consistent with prior findings in speech emotion recognition, where vocal cues such as pitch variation and intensity strongly convey emotions like anger and sadness. The gap between unimodal and multimodal performance underscores that neither modality alone is sufficient for robust emotion recognition, especially under subject-independent conditions.

3.3. Ablation Study and Fusion Strategy Analysis

The ablation study was designed to disentangle the individual contributions of modality selection, temporal attention, and cross-modal fusion within the proposed framework. Table 4 summarizes the performance of audio-only (A), video-only (V), and combined audio–visual (AV) configurations, evaluated with and without attention mechanisms.

Table 4. Performance comparison of unimodal and multimodal simple concatenation configurations with and without attention.

Dataset	Attention	Audio (A)	Video (V)	Audio–Video (AV)
RAVDESS	Yes	68.2	59.7	**78.9**
RAVDESS	No	64.7	58.0	71.6
CREMA-D	Yes	59.5	56.7	–
CREMA-D	No	57.3	53.1	–

The results indicate that audio-only models consistently outperform video-only models across both datasets, reflecting the strong affective information encoded in speech prosody. Introducing temporal attention improves performance for both modalities, confirming that selectively weighting emotionally salient time steps enhances sequence-level representation. The confusion matrix in Figure 7 illustrates that When audio and visual modalities are combined using simple feature concatenation, performance improves substantially relative to unimodal baselines.

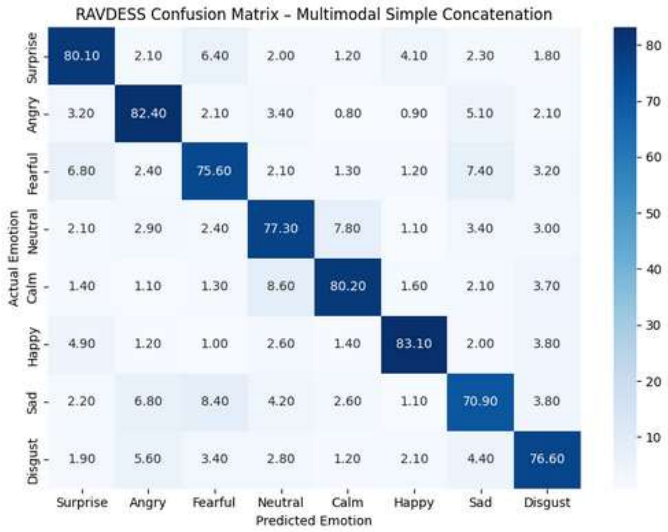

Figure 7. Confusion Matrix of multimodal simple concatenation on the RAVDESS Dataset

On RAVDESS, this configuration achieves an accuracy of 78.9%, representing an improvement of over 10% compared to the best unimodal result. This gain confirms that multimodal integration is beneficial even under a relatively simple fusion strategy, as complementary affective cues from facial expressions and vocal signals jointly contribute to emotion discrimination.

However, simple concatenation remains limited in its ability to fully exploit multimodal interactions. By treating audio and visual features uniformly across time, it does not explicitly account for temporal misalignment between modalities or varying modality reliability under different emotional expressions. As a result, informative cues may be diluted by redundant or noisy features, particularly for emotions with subtle or asynchronous expression patterns.

The full proposed framework addresses these limitations by integrating temporal attention within each modality and cross-modal attention at the fusion stage. Unlike simple concatenation, cross-modal attention explicitly aligns temporally encoded audio and visual representations, enabling the model to dynamically emphasize the modality that carries the most informative cues at each moment. This structured interaction leads to a substantial performance gain, with the complete framework achieving 91.22% accuracy on RAVDESS and 87.32% on CREMA-D.

These results demonstrate that the performance improvement is not merely a consequence of combining modalities, but rather of how temporal modeling and cross-modal alignment are jointly

enforced. The ablation study therefore confirms that temporal attention and cross-modal attention play complementary roles: temporal attention refines intra-modal representations, while cross-modal attention enables effective synchronization and weighting of audio–visual cues prior to classification.

3.4. Discussion and Implications

The experimental findings demonstrate that robust emotion recognition emerges from the joint modeling of deep spatial features, temporal dynamics, and adaptive multimodal fusion. The integration of ResNet-50 and BiLSTM enables effective capture of both instantaneous expressive cues and longer-term emotional trajectories, while attention-driven fusion facilitates coherent interaction between modalities. The reduced confusion between acoustically similar or visually subtle emotion pairs can be attributed to the cross-modal attention mechanism, which enables the model to resolve ambiguity by selectively reinforcing complementary cues across modalities.

The consistent gains observed under subject-disjoint evaluation highlight the framework's strong generalization capability, addressing a major limitation of many prior studies. These properties make the proposed approach well suited for real-world applications where speaker variability, environmental noise, and expressive diversity are unavoidable.

4. CONCLUSIONS

This work introduced a fully integrated deep learning framework for temporal multimodal emotion recognition that unifies deep spatial representation, sequence-level temporal modeling, and feature-level cross-modal fusion within a single end-to-end architecture. Unlike prior approaches that treat temporal modeling and multimodal fusion as loosely coupled components, the proposed framework applies a consistent temporal learning strategy across both audio and visual streams using BiLSTM networks, followed by explicit cross-modal alignment through attention-driven fusion. This design enables coherent modeling of asynchronous and complementary affective cues.

Extensive evaluation on the RAVDESS and CREMA-D benchmarks under strict subject-disjoint protocols demonstrates that the proposed approach consistently outperforms unimodal baselines and recent multimodal methods. Detailed class-wise analyses and confusion-matrix-based evaluations further reveal that attention-guided fusion substantially improves discrimination for emotionally ambiguous and overlapping categories, confirming the effectiveness of aligning modality-specific temporal representations prior to classification. The stability of performance across folds highlights the robustness and generalization capability of the framework under realistic variability in speakers and expressive styles.

Overall, the study advances multimodal affective computing by demonstrating that robust emotion recognition emerges from the joint optimization of deep spatial encoding, temporal dependency modeling, and adaptive cross-modal interaction. The proposed framework provides a principled and extensible foundation for future research in multimodal emotion analysis, with clear pathways for extending the architecture to additional modalities, longer temporal contexts, and real-time affect-aware systems operating under unconstrained conditions.

5.0 LIMITATIONS, PRACTICAL CONSIDERATIONS, AND ETHICAL IMPLICATIONS

Despite the consistent performance gains achieved by the proposed framework, several limitations should be acknowledged. The model assumes reasonably reliable face detection and audio quality; under severe occlusion, extreme head pose, or high background noise, one or both modalities may be degraded, leading to residual confusion between closely related emotional categories such as neutral and calm. From a computational perspective, the use of parallel deep residual backbones, BiLSTM-

based temporal modeling, and attention-guided fusion increases training cost relative to unimodal systems, making the framework more suitable for offline training or near–real-time deployment in resource-equipped environments. Once trained, however, inference remains efficient due to the feed-forward nature of feature extraction, temporal modeling, and fusion, supporting practical use in controlled applications such as e-learning and human–computer interaction systems.

Audio–visual emotion recognition also raises important ethical considerations, as facial expressions and vocal characteristics constitute sensitive biometric data that require explicit user consent, transparent data governance, and secure handling to protect privacy [31]. In addition, models trained on datasets with limited demographic balance may exhibit biased performance across populations, a concern widely documented in facial analysis research [32]. Consequently, emotion predictions should be treated as supportive signals rather than definitive judgments, with responsible deployment emphasizing transparency, contextual awareness, and human oversight.

Acknowledgements

The authors acknowledge the support provided by Africa Centre of Excellence on Technology Enhanced Learning, National Open University of Nigeria toward this study. The first author conducted all experiments and prepared the initial draft of the manuscript under the supervision, guidance, and contributions of the co-authors.

References

[1] E. Ghaleb, J. Niehues, and and S. Asteriadis, “MULTIMODAL ATTENTION-MECHANISM FOR TEMPORAL EMOTION RECOGNITION,” *IEEE Int. Conf. Image Process.*, pp. 251–255, 2020.

[2] H. M. Shahzad, S. M. Bhatti, A. Jaffar, and M. Rashid, “A Multi-Modal Deep Learning Approach for Emotion Recognition,” *Intell. Autom. Soft Comput.*, vol. 36, no. 2, pp. 1561–1570, 2023, doi: 10.32604/iasc.2023.032525.

[3] R. E. Donatus, U. O. Chiedu, and I. H. Donatus, “Exploring the Impact of Convolutional Neural Networks on Facial Emotion Detection and Recognition,” *Asian J. Electr. Sci.*, vol. 13, no. 1, pp. 35–45, 2024.

[4] Y. Fu, Q. Liu, Q. Song, P. Zhang, and G. Liao, “Multi-HM: A Chinese Multimodal Dataset and Fusion Framework for Emotion Recognition in Human–Machine Dialogue Systems,” *Appl. Sci.*, vol. 15, no. 8, p. 4509, 2025.

[5] S. Kalateh, L. A. Estrada-Jimenez, S. Nikghadam-Hojjati, and J. Barata, “A systematic review on multimodal emotion recognition: building blocks, current state, applications, and challenges,” *IEEE Access*, vol. 12, pp. 103976–104019, 2024.

[6] P. Srinivas and P. Mishra, “Human Emotion Recognition by Integrating Facial and Speech Features: An Implementation of Multimodal Framework using CNN,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 1, 2022.

[7] S. Khuntia, A. Amjad, R. B. Tarekegen, and L.-C. Tai, “Deep Learning-Based Emotion Recognition Using Fusion of Multimodal Affective Data From Consumer-Grade Wearable ECG and Speech Sensors,” in *2025 IEEE International Conference on Consumer Electronics (ICCE)*, IEEE, 2025, pp. 1–6.

[8] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, “A survey of deep learning-based multimodal emotion recognition: Speech, text, and face,” *Entropy*, vol. 25, no. 10, p. 1440, 2023.

[9] S. Zhang, Y. Yang, C. Chen, X. Zhang, Q. Leng, and X. Zhao, “Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects,” *Expert Syst. Appl.*, vol. 237, p. 121692, 2024.

[10] M.-H. Yi, K.-C. Kwak, and J.-H. Shin, "HyFusER: hybrid multimodal transformer for emotion recognition using dual cross modal attention," *Appl. Sci.*, vol. 15, no. 3, p. 1053, 2025.

[11] G. Udahemuka, K. Djouani, and A. M. Kurien, "Multimodal Emotion Recognition Using Visual, Vocal and Physiological Signals: A Review," *Appl. Sci.*, vol. 14, no. 17, 2024, doi: 10.3390/app14178071.

[12] M. K. Chowdary, T. N. Nguyen, and D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications," *Neural Comput. Appl.*, vol. 35, no. 32, pp. 23311–23328, 2023.

[13] N.-H. Ho, H.-J. Yang, S.-H. Kim, and G. Lee, "Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network," *IEEE Access*, vol. 8, pp. 61672–61686, 2020.

[14] P. Varshney, R. Dey, V. Gulati, and D. K. Vishwakarma, "Speech Emotion Recognition: A Multimodal Approach Using Multi-Feature Fusion and Self-Attention," in *2025 6th International Conference on Inventive Research in Computing Applications (ICIRCA)*, IEEE, 2025, pp. 1056–1063.

[15] R. E. Donatus, "Interpretable Speech Emotion Recognition : A Comparative Study of BiLSTM Temporal Attention and Transformer-Based," *Asian J. Electr. Sci.*, vol. 14, no. 2, pp. 21–27, 2025.

[16] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Inf. fusion*, vol. 37, pp. 98–125, 2017.

[17] A. I. Middya, B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowledge-Based Syst.*, vol. 244, p. 108580, 2022, doi: 10.1016/j.knosys.2022.108580.

[18] M. Khan, W. Gueaieb, A. El Saddik, and S. Kwon, "MSER: Multimodal speech emotion recognition using cross-attention with deep fusion," *Expert Syst. Appl.*, vol. 245, p. 122946, 2024.

[19] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Trans. circuits Syst. video Technol.*, vol. 28, no. 10, pp. 3030–3043, 2017.

[20] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Inf. Fusion*, vol. 49, pp. 69–78, 2019.

[21] Z. Farhoudi and S. Setayeshi, "Fusion of deep learning features with mixture of brain emotional learning for audio-visual emotion recognition," *Speech Commun.*, vol. 127, no. June 2020, pp. 92–103, 2021, doi: 10.1016/j.specom.2020.12.001.

[22] D. E. Rexcharles, B. L. Pal, I. H. Donatus, and U. O. Chiedu, "Comparative Analysis of Spectrogram and MFCC Representations for Speech Emotion Recognition Using Machine Learning," *Asian J. Comput. Sci. Technol.*, vol. 13, no. 2, pp. 41–47, 2024.

[23] V. Gupta, S. Juyal, G. P. Singh, C. Killa, and N. Gupta, "Emotion recognition of audio/speech data using deep learning approaches," *J. Inf. Optim. Sci.*, vol. 41, no. 6, pp. 1309–1317, 2020.

[24] L. Schoneveld, A. Othmani, and H. Abdelkawy, "Leveraging recent advances in deep learning for audio-visual emotion recognition," *Pattern Recognit. Lett.*, vol. 146, pp. 1–7, 2021.

[25] B. Mocanu, R. Tapu, and T. Zaharia, "Multimodal Emotion Recognition using Cross Modal Audio-Video Fusion with Attention and Deep Metric Learning," *Image Vis. Comput.*, vol. 133, pp. 1–18, 2023.

[26] S. S. Poorna, V. Menon, and S. Gopalan, "Hybrid CNN-BiLSTM architecture with multiple attention mechanisms to enhance speech emotion recognition," *Biomed. Signal Process. Control*, vol. 100, p. 106967, 2025.

[27] W. Fan, X. Xu, G. Zhou, X. Deng, and X. Xing, "Coordination Attention based Transformers with bidirectional contrastive loss for multimodal speech emotion recognition," *Speech Commun.*, vol. 169, p. 103198, 2025.

[28] F. Liu, Z. Fu, Y. Wang, and Q. Zheng, "TACFN: transformer-based adaptive cross-modal fusion network for multimodal emotion recognition," *arXiv Prepr. arXiv2505.06536*, 2025.

[29] M. Aly, "Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model," *Multimed. Tools Appl.*, vol. 84, no. 13, pp. 12575–

12614, 2025.

[30] R. G. Praveen, E. Granger, and P. Cardinal, "Cross attentional audio-visual fusion for dimensional emotion recognition," *16th IEEE Int. Conf. Autom. face gesture Recognit.*, pp. 1–8, 2021.

[31] D. Barker, M. K. R. Tippireddy, A. Farhan, and B. Ahmed, "Ethical Considerations in Emotion Recognition Research," *Psychol. Int.*, vol. 7, no. 2, p. 43, 2025.

[32] L. Goncalves and C. Busso, "Robust audiovisual emotion recognition: Aligning modalities, capturing temporal information, and handling missing features," *IEEE Trans. Affect. Comput.*, vol. 13, no. 4, pp. 2156–2170, 2022.