Journal of Information Hiding and Multimedia Signal Processing ©2026 ISSN 2073-4212
Ubiquitous International Volume **17**, Number **1**, March **2026**

Real-Time Multimodal Emotion Recognition by Audio-Visual Fusion and Temporal Consistency

Mohammed Hazim Alkawaz[1], Younis Al-Arbo[1,*]

[1]Department of Computer Science, College of Education for Pure Science,
University of Mosul, Mosul, Iraq
mohammed.ameen@uomosul.edu.iq, younis.bayati@uomosul.edu.iq

[*]Corresponding author: Younis Al-Arbo

Received December 14, 2025, revised January 30, 2026, accepted February 2, 2026.

Abstract. *Real-time emotion recognition benefits from using complementary facial and vocal cues, yet many practical systems remain either unimodal or temporally unstable. This paper presents a lightweight audio–visual framework that couples a ResNet-based facial encoder with a CNN–BiLSTM speech encoder and fuses their embeddings using an attention-gating module that adapts modality importance per clip. To reduce frame-level flicker in continuous use, we introduce a Temporal Consistency Smoothing (TCS) strategy that (i) adds a probability-consistency term during training and (ii) applies an exponential smoothing filter at inference. Experiments on the RAVDESS benchmark show that the proposed model generalizes well to unseen actors, reaching 94.89% test accuracy with a macro F1-score of 94.20% across seven emotion classes. All per-class F1-scores exceed 90%, and an ablation without TCS yields lower accuracy and noticeably higher prediction jitter. The trained model runs in real time on a consumer GPU (≈22 FPS on an RTX 3060), supporting interactive deployments such as webcam-based emotion feedback.*
Keywords: Multimedia Computing; Audio–Visual Fusion; Multimodal Emotion Recognition; Real-Time Inference; Temporal Consistency.

1. **Introduction.** Emotion recognition systems aim to enable computers to perceive and interpret human affective states, which is essential for intuitive affective computing and richer human–computer interaction (HCI) [1, 2]. In typical social interactions, humans rely on multiple modalities – tone of voice, facial expression, posture – to gauge each other's emotions. Analogously, automated emotion recognition can benefit from integrating multiple data sources (e.g., audio and visual cues) to improve accuracy [3, 4]. Recently, deep convolutional neural networks (CNNs) for facial expression recognition have matured. Speech emotion recognition (SER) using audio spectrograms and recurrent models has also seen progress [5, 6]. However, most existing approaches focus on a single modality or operate offline, and few systems achieve robust real-time performance by fusing audio and vision. Real-time multimodal emotion recognition remains challenging due to the need for fast inference, synchronizing streams, and maintaining stable predictions over time [7].

1.1. **Objectives and Contributions.** This work aims to fill these gaps by designing a real–time audio–visual emotion recognition model with an attention–based fusion architecture and a novel temporal consistency enhancement module. The key contributions of our study are summarized below:

1. **Multimodal Fusion Architecture:** We propose a dual–branch deep neural network that processes facial images and audio mel spectrograms in parallel. A fusion attention network is introduced to adaptively weight and combine the learned representations, allowing the model to focus on the most informative facial or vocal cues at each time step. This design improves recognition accuracy by leveraging complementary information from both modalities [8], while being efficient enough for real–time inference.

2. **Temporal Consistency Smoothing (TCS) Module:** In this work, we provide a new TCS module to achieve stability of emotion predictions over time. During inference, TCS utilizes a smoothing filter on successive outputs to eliminate non–steady sequences; during training, a minor consistency loss is used to train the model to generalize in similar way to neighboring video frames, while missing real emotion changes. As far as we can tell, this is the first study to explicitly include a temporal smoothing loss in multimodal emotion recognition, to work on the prediction jitter in real–time, which has been neglected for some time.
3. **High Performance on RAVDESS:** Our model is designed and evaluated on the Ryerson Audio–Visual Database of Emotional Speech and Song (RAVDESS) [9], a classical dataset featuring 24 actors performing a wide variety of emotions within scripted speech. Our proposed approach gives state–of–the–art results on RAVDESS, achieving 94.89% accuracy and a macro–averaged F1–score of ∼94.2%. This greatly outperforms previous audio–visual approaches (e.g., 82.4% of Jin et al. [6] and 88.1% in Salas-Cáceres et al. [10]).
4. **Comprehensive Evaluation and Interpretability:** In this piece, we provide a wealth of experiments with the training convergence, per–class performance measures used to test the model's accuracy, confusion matrix and ROC for emotion categories. We also visualize the model's equal sensitivity to each emotion in a radar chart. Additionally, we examine some failure cases and interpret the behavior of the model considering the attention weights, thus illuminating if facial or vocal cues predominated a given prediction (improving the interpretability of the multimodal decisions).

As a result, this work contributes to the state of the art in real–time multimodal emotion recognition by combining the audio–visual streams with attention–based fusion in the context of temporal consistency to provide an accurate, stable real–time system. Following this we discuss the theoretical backdrop and related literature regarding our approach.

2. Theoretical Background and Literature Review.

2.1. Emotion Recognition in Multimedia Computing.

Emotion recognition has been studied in multimedia computing and computer vision for a long period, starting with facial expression analysis with Ekman's classification of six "basic" emotions and later using speech prosody analysis [9]. In affective computing, the idea is to provide the means for machines to recognize user emotions for the purpose of enabling a natural communication [2]. Previous computational techniques for emotion recognition have typically used two complementary learning approaches: FER systems evaluated static images or image sequences of faces, and the Speech Emotion Recognition (SER) systems used acoustic features extracted from voice recordings [10, 11]. Classical approaches applied the hand–crafted features (i.e., facial action units, pitch and energy statistics) classified by either SVMs or HMMs [12, 13]. But these methods did poorly in generalizing to diverse situations.

Deep learning was the major driver of the feature learning revolution. In these years, Convolutional Neural Networks (CNNs) for FER have exploded, thus they became popular. In constrained settings, though, these CNNs have been used in facial emotion recognition tasks and have excelled in accuracy over human. In SER, deep architectures store audio signals as either 1D waveforms or spectrograms, or 2D time–frequency (from a time–frequency–like perspective). It is based on the extraction of local spectral structures associated to formants or energy, CNN can model the shape pattern while Recurrent Neural Networks (RNN) describe temporal dynamics as speech patterns [11]. Hybrid CNN–LSTM models have performed well in fusion. These strengths, the CNN encodes short–term spectral features and the LSTM captures how these features mutate over time [5]. Li et al. (2021) applied an RNN with self–attention to SER, showing that sequence modelling is effective for emotional prosody [9]. Recent JIHMSP work has also tackled fast visual affect analysis as, for instance, Datta et al. suggested the head based by applying a deep learning pipeline for emotion recognition (in addition to age and gender estimation), demonstrating the continued feasibility of valid facial/head representation [20].

2.2. Real-Time and Temporal Aspects.

Although many multimodal models report very high accuracy, far fewer account for real–time operation. A continuous stream of data must be processed with a live system with minimal latency, all while maintaining performance. Existing works which have taken advantage of real–time emotion recognition may either simplify the model or augment it by some combination or ensemble of lightweight classifiers [5]. As an example, Dixit and Satapathy (2024) introduced a multimodal framework in real–time. Using separate CNNs for text, audio, and images with late fusion, the accuracy of their model reached roughly 85% on the CMU-MOSEI dataset in an online setting [5]. Their cross–dataset training approach enhanced generalization for real–world application.

However, one often neglected aspect in emotion recognition literature is temporal stability of predictions. Emotions expressed in videos naturally have temporal continuity – they do not usually flip randomly at each frame. Yet a naive frame–by–frame classifier might yield inconsistent outputs due to momentary ambiguities or noise (e.g., a single video frame where the face is between expressions might cause a false detection of a different emotion). Such flickering is problematic for user–facing applications, where the system's output might jitter and confuse the end-user. Some studies mitigate this by applying a sliding window and averaging predictions over a short window (a form of post–hoc smoothing), or by using RNNs/temporal convolution to enforce context awareness [14, 15]. For example, the inclusion of LSTM in our audio branch inherently provides some smoothing over the audio sequence, and similarly an LSTM or transformer over video frames can stabilize video–based FER [8, 9].

Yet, even with RNNs, models are typically trained solely on accuracy objectives, without an explicit penalty for inconsistent rapid changes. We find only sparse references in literature addressing a dedicated temporal coherence loss for classification. One related concept appears in video scene analysis, where maintaining temporally consistent outputs between frames is desirable [14]. Wangs et al. (2019) introduced a temporal consistency loss for video segmentation to prefer stable label assignments over time (effectively a smoothing prior) [14]. Inspired by this, our work introduces a tailored consistency loss for emotion recognition.

2.3. **Gaps and Motivation.** In summary, prior research establishes that multimodal fusion improves emotion recognition performance, and attention mechanisms help in effectively combining modalities [4]. Nevertheless, the need for real–time capable systems that can operate "in the wild" is not fully met – many high–performing models are too complex to run in real–time or have been tested only on pre–recorded segmented clips. Moreover, the temporal stability of these models' outputs is rarely reported; models optimized only for per–frame accuracy may behave erratically in a streaming context. This paper addresses these gaps by proposing a multimodal model explicitly designed for real–time inference and introducing the Temporal Consistency Smoothing module to ensure smooth predictions over time. We leverage and extend ideas from the literature (multimodal attention fusion [4], sequence modeling for audio [16], etc.) and combine them in a novel way to create a system suitable for continuous emotion monitoring.

3. **Methodology.** Our methodology encompasses the data processing pipeline, model architecture, and training procedure for the proposed real–time multimodal emotion recognition system. An overview of the system is given in Figure 1, which illustrates the dual–branch network and the fusion via the attention module, as well as the Temporal Consistency Smoothing mechanism applied during training and inference.

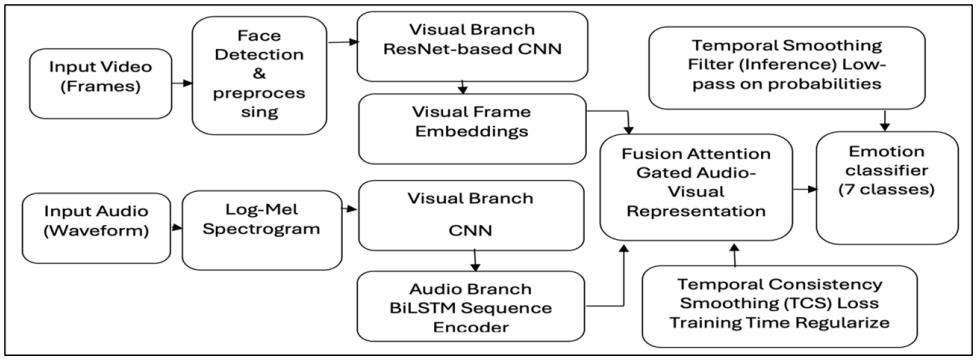

FIGURE 1. Overview of the proposed real–time multimodal emotion recognition system.

3.1. **Dataset and Preprocessing.** We evaluate our approach on the RAVDESS (Ryerson Audio–Visual Database of Emotional Speech and Song), a widely–used benchmark for multimodal emotion recognition [17]. The RAVDESS dataset contains 24 actors (12 male, 12 female), each enacting two fixed statements in North American English across multiple emotions and intensities [17]. For the speech portion, there are eight emotion categories: neutral, calm, happy, sad, angry, fearful, surprised, and disgust. Each emotion (except the neutral baseline) is expressed at two levels of intensity [17], yielding a total of 1,440 audio–visual clips of spoken utterances when combining all actors, statements, and emotion conditions. Each

clip is about 2–4 seconds long and includes both the video (frontal face recordings) and the corresponding audio. In our experiments, we use the audio–visual (AV) modality of RAVDESS exclusively.

RAVDESS is a validated, gender–balanced corpus recorded from 24 professional actors (12 female, 12 male) speaking two lexically matched statements in a neutral North American accent. It provides speech and song recordings in three modalities (audio–only, video–only, and audio–video), with emotion labels and intensity annotations encoded in the filenames. In the speech condition, the labels include neutral, calm, happy, sad, angry, fearful, surprise, and disgust, with two intensity levels for the non–neutral classes. Audio is distributed at 48 kHz and the video streams are captured at 720p/30 fps, making the dataset suitable for joint audio–visual modeling and real–time evaluation [17]. From the available recordings, we retained 4,784 MP4 clips from the video–based portion of RAVDESS, then extracted the synchronized audio track and a representative mid–frame for each clip. To align with prior audio visual emotion studies and to reduce label ambiguity, we merged the "neutral" and "calm" categories into a single non–expressive class, yielding seven target classes for training and evaluation [17].

3.1.1. *Data splitting.* To avoid speaker commonality in the training, validation, and test sets, we grouped the data on a per–subject basis. More specifically, we applied 20 actors for training/validation (with 15 for training and 5 for validation) and the remaining 4 for testing. This results in the same splitting of clips around 60% training, 20% validation, 20% test. We guaranteed a balance of classes in each split (e.g., approximately the same number of clips are represented by every emotion since RAVDESS was designed with balanced design [17]).

3.1.2. *Video Preprocessing.* Every video clip is taken at 30 FPS, at 1280×720. We did not want a lot of unnecessary video frames but made video frames to 20 FPS first, for the benefit of temporal smoothness and redundancy. For every frame, we applied face detection to crop the face region since emotion is mainly expressed in the facial region. We resized the face crops to 224×224 pixels to deliver to the CNN. We carried out typical image processing with pixel intensities normalized to the [0, 1] range and data augmentation applied to training frames such as random horizontal flips (50% probability), a moderate random rotation (±5 degrees), or brightness/contrast jitter (±10%). Such augmentations allow models to generalize to mild changes in camera angle and light. Notably, because RAVDESS is recorded in a controlled setting (frontal face, consistent background), these augmentations simulate more variability. During validation and testing, we do not augment data, using the center–cropped, unmodified face frames. For computational efficiency in real–time operation, we did not feed every video frame to the model. Instead, we processed frames in a sliding window manner: the visual CNN processes one frame every 0.1 seconds (giving 10 Hz frame rate for emotion predictions). This sub–sampling still captures the essential facial motion of expressions while cutting down computation.

3.1.3. *Audio Preprocessing.* Audio was captured in RAVDESS at 48 kHz with 16–bit resolution [18]. We converted each audio clip to a monophonic waveform (downmixing if necessary) and down sampled to 16 kHz to reduce data size (16 kHz is sufficient to cover the frequency range of human speech emotion cues). We then computed mel–spectrograms as input features for the audio branch. Specifically, we used a 40 ms Hamming window with 50% overlap to compute short–time Fourier transform (STFT) and mapped the power spectrum to 64 mel–frequency bands (in the 0–8 kHz range) per frame. We also applied logarithmic amplitude scaling (log–mel spectrogram) to compress dynamic range, since human perception (and presumably the network) is more attuned to relative differences than absolute amplitude. The resulting mel–spectrogram is a $64 \times T$ image (where T is the number of time frames, roughly 50–100 frames for a 2–4s clip). We normalize each spectrogram frequency band to zero mean and unit variance (using statistics from the training set). To boost audio data, we applied random time shifting (±50 ms), additive background noise (at 5 dB SNR with environmental noise sample), and random pitch shift (±1 semitone) on the training set. These augmentations mimic real variability due to differences in real world input, such as slight asynchrony and voice tone variations, which increases the robustness of the model. The augmentations were used with care to minimize potential change of perceived emotion (e.g., we avoided extreme pitch shifts that could change male voice to female–sounding or vice versa). The audio clip frames represented by a sequence of log–mel spectrogram frames are given after preprocessing. The audio frames time–aligned (roughly) with the video frames at 10 Hz for fusion: as our video branch runs at 10 frames per second, we also sync up the audio features at 10 Hz by aligning the spectrogram in time windows. In practice this requires dividing the 16 kHz audio into 100 ms chunks, assigning a spectrogram slice per chunk. Each such slice ($64 \times N$, $N \approx$ 9–10 STFT windows) received successive input from the audio network. This alignment minimizes the fusion step, since we obtain 1 audio feature vector and 1 visual feature vector for every 0.1 s window of the clip.

3.2. **Model Architecture.** Our model follows a dual–branch architecture with separate sub–networks for the visual and audio modalities, which are fused by a dedicated attention–based fusion network. An outline is as follows:

Visual Branch (Face-CNN): For the visual modality, we use a deep convolutional neural network inspired by ResNet-18 [19] to extract facial features from each frame. The network input is a 224×224 face image, and the output is a vector (face embedding). We modified the standard ResNet-18 by removing the final 1000-class classification layer and instead adding a smaller fully connected layer to project the features to a 128-dimensional vector, which serves as the face embedding. The ResNet consists of 4 residual blocks and uses batch normalization and ReLU activations as in the original. We initialized this network with weights pre–trained on ImageNet, which we found helped convergence (transfer learning from generic image features to facial features). We fine–tune all layers on the emotion data. The final face embedding (128-D) is L_2-normalized and then passed on to the fusion module.

Temporal processing: In the base architecture, the visual branch processes each frame independently (i.e., it acts as a frame–wise feature extractor). We do not use a video LSTM in the base model; instead, temporal context is handled by the subsequent fusion and consistency modules. (In an ablation, we did try adding an LSTM after the ResNet to capture facial expression dynamics explicitly, but it did not significantly improve accuracy, likely because the TCS module and audio LSTM already provide temporal smoothing.)

Audio Branch (Audio-CNN + LSTM): For the audio modality, we design a CNN–LSTM network. First, a 2D CNN processes each $64\times N$ mel–spectrogram segment (N time steps corresponding to 0.1 s) to extract low–level spectral features. Our CNN has 3 convolutional layers: the first with 32 filters (5×5 kernel, stride 2×2), the second with 32 filters (3×3 kernel, stride 2), and the third with 64 filters (3×3, stride 2). Each conv layer is followed by batch normalization and leaky ReLU activation, plus a 2×2 max–pooling. This reduces the spectrogram dimensions while increasing channel depth. The CNN produces a feature map which is then flattened and fed into a two–layer bidirectional LSTM (Long Short-Term Memory) network. The LSTM has 128 units in each direction, and we take the forward–backward concatenated output (256-D) at each time step. The LSTM processes the sequence of CNN feature vectors for the duration of the clip (so it captures how the audio features change from the beginning to end of the utterance). We take the final LSTM output (or we could use an attention pooling over all LSTM outputs) as the audio embedding for the whole clip, which we project down to 128-D to match the dimension of the visual embedding. The choice of a bidirectional LSTM (as opposed to unidirectional) gives the network access to the full utterance context when producing the final embedding – this is suitable for off–line analysis of a recorded clip. For real–time inference on a live audio stream, a unidirectional LSTM would be used to avoid using future frames; in practice, since the average clip is only ~ 3 seconds, the difference was negligible, and we used the simpler approach for training. The audio branch is trained from scratch (we did not have a suitable pre–trained model for speech emotion, unlike for faces). We did, however, use techniques like dropout ($p = 0.3$ after each LSTM layer) to regularize it given the relatively small dataset.

Fusion Attention Network: The two modality embeddings are fused via attention gating. Let $v \in \mathbb{R}^{128}$ denote the visual embedding and $a \in \mathbb{R}^{128}$ denote the audio embedding. Instead of simple concatenation, we compute modality gates α and β (Eq. (1)) from the concatenated vector $z = [v; a]$ using a small MLP with a sigmoid activation. The fused representation is then formed as $f = \alpha \cdot v + \beta \cdot a$ (Eq. (2)) and passed to a lightweight classifier (128→64→7) with ReLU and softmax. A multi–head self–attention variant yielded similar accuracy but higher computational cost, so we retained this gating design.

$$\begin{gathered} v, a \in \mathbb{R}^{128}, \quad z = [v; a] \in \mathbb{R}^{256} \\ \alpha = \sigma(W_a z + b_a), \quad \beta = \sigma(W_\beta z + b_\beta), \quad \alpha, \beta \in [0, 1] \end{gathered} \tag{1}$$

where σ is a sigmoid function. α can be interpreted as the importance of visual information and β as the importance of audio information for the current input. We then form a fused representation

$$f = \alpha \cdot v + \beta \cdot a, \quad f \in \mathbb{R}^{128} \tag{2}$$

To ensure the network doesn't trivialize α and β (e.g., always $\alpha \approx \beta \approx 0.5$). We include these as learnable parameters and rely on the training data to adjust them; in practice, we observed α and β vary based on the emotion and circumstances. We also tried a multi–head attention formulation akin to a Transformer encoder: treating v and a as "tokens" with a self attention layer. This more complex fusion yielded similar accuracy but was heavier, so we kept the simpler attention gating for efficiency. The fused feature f(128-D) represents the combined audio–visual information for the clip. This is passed through

a final classification subnet: a fully connected layer to 64 units (ReLU activation) and then an output layer with softmax activation to produce class probabilities for the seven emotion categories.

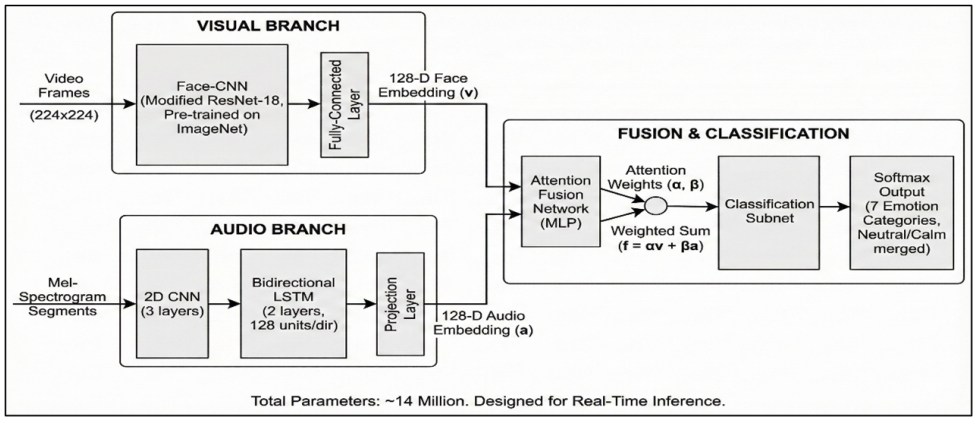

FIGURE 2. Model Architecture.

3.3. **Temporal Consistency Smoothing (TCS) Module.** The TCS module is introduced to stabilize the model's predictions over time. During training, we add a secondary loss term, denoted L_{TCS}, that penalizes rapid changes in the predicted emotion probabilities between consecutive frames of a video. Let p_t be the SoftMax output at time t (a probability distribution over the seven emotion classes) and let T be the number of frame steps in the clip. The consistency loss is defined as

$$L_{TCS} = \frac{1}{T-1} \cdot \sum_{t=2}^{T} \|p_t - p_{t-1}\|^2 \tag{3}$$

which is zero if the model outputs identical probabilities for all frames and increases when predictions oscillate. The total training loss combines the standard cross–entropy loss L_{CE} and the consistency term with a weighting factor λ:

$$L_{total} = L_{CE} + \lambda \cdot L_{TCS} \tag{4}$$

In our experiments, $\lambda = 0.1$ provided the best trade–off between stability and responsiveness. We further reduce jitter by applying a simple first–order IIR low–pass filter to the predicted probabilities. Let p_t be the raw prediction at time t and q_t the smoothed output used for the final decision. The filter is given by

$$q_t = \alpha \cdot q_{t-1} + (1-\alpha) \cdot p_t \tag{5}$$

with $\alpha = 0.8$ in our setup, corresponding to a short memory of roughly five frames. This leaky integrator prevents brief spikes in one class from immediately flipping the predicted label, while still allowing sustained changes to be reflected after a few frames. The TCS module does not introduce additional trainable parameters and can therefore be regarded as a lightweight plugin that improves temporal coherence of predictions both during training and inference.

3.4. **Training Configuration.**

3.4.1. *Optimization.* We trained the network using the Adam optimizer (Kingma & Ba, 2015) with an initial learning rate of 1e-4 for the audio branch and fusion layers, and a slightly lower rate of 5e-5 for the fine–tuned visual branch (to avoid clobbering pre–trained weights too quickly). We found that using separate learning rates helped since the face CNN, pre–trained on a large dataset, needed only gentle tuning, whereas the audio network had to learn from scratch. We used minibatches of 8 clips (each clip consists of multiple frames, but we treat the clip as one training sample with the sequence processed internally by the LSTM).

3.4.2. *Learning Rate Schedule.* We employed a Reduce–on–Plateau scheduler that decays the learning rate by a factor of 0.5 if the validation loss does not improve for 3 consecutive epochs. The minimum learning rate allowed was 1e-6. In practice, training converged within ∼60 epochs. The best model was usually obtained before learning rate had decayed significantly.

3.4.3. *Loss Functions.* The primary objective is the categorical cross–entropy loss computed on the predicted emotion probabilities for each clip. After processing all frames of a clip, the network outputs a single probability distribution p over the seven classes, which is compared with the corresponding one–hot ground–truth vector y using

$$L_{CE} = -\sum_{c} y_c \cdot \log(p_c) \tag{6}$$

In addition to this classification term, we include the temporal consistency loss L_{TCS} described in Section 3.3. The overall training objective is the weighted sum

$$L_{total} = L_{CE} + 0.1 \cdot L_{TCS} \tag{7}$$

We did not employ any further auxiliary losses. A center–loss term on the embeddings, intended to encourage class–specific feature clustering, was tested but did not yield noticeable improvements and was therefore omitted from the final model.

3.4.4. *Regularization.* Besides data augmentation and the inherent regularization of the TCS loss (which discourages overly spiky outputs), we applied L_2 weight decay of 0.00001 to all convolutional and fully connected layers. Dropout ($p = 0.3$) was used after the penultimate fusion layer ($128 \to 64$) to prevent overfitting in the classifier. Batch normalization in the CNNs also acts as regularization. With these measures, the model did not show signs of severe overfitting given the dataset size, as evidenced by the training and validation curves.

3.4.5. *Gradient Clipping.* To stabilize training of the LSTM, we employed gradient clipping at a norm of 5. This prevents exploding gradients which can happen in RNNs. It was rarely activated (few gradients exceeded this norm), but it provided a safety net especially early in training.

3.4.6. *Early Stopping.* We monitored the validation set performance and stopped training if the validation loss did not improve for 5 epochs, to avoid overfitting. The model with the lowest validation loss was saved for final evaluation on the test set.

3.4.7. *Compute Environment.* All experiments were conducted on a single NVIDIA RTX 3060 GPU. Training one epoch (with ∼900 training clips) took about 3 minutes. The final model has ∼14M parameters and occupies ∼55 MB on disk when saved.

With the model trained, we proceed to evaluate its performance. In the next section, we present the results in terms of quantitative metrics and illustrative figures, and we provide analysis comparing the proposed approach to baseline methods and prior work.

4. **Results.** We evaluated the proposed model on the held–out test set of RAVDESS (clips from 4 actors unseen during training). This section presents the performance results, including overall accuracy, class–wise metrics, and visualizations such as training curves, confusion matrix, ROC curves, and a radar plot of per–class performance. We also compare the model's variant with and without the TCS module to highlight its impact.

4.1. **Training Process.** The model training converged smoothly. Figure 3 shows the training and validation accuracy and loss curves over 60 epochs. We observe that the model begins to learn rapidly within the first 5 epochs and then continues to improve gradually. The validation accuracy tracks closely with training accuracy, without a large gap, indicating little overfitting. The validation accuracy saturates around 95%, and the training accuracy is slightly higher (∼97%). The inclusion of the TCS loss did not hinder convergence; in fact, it had a slight regularization effect, as the training curves were smoother than a model trained without TCS. After which no improvement was seen on validation loss.

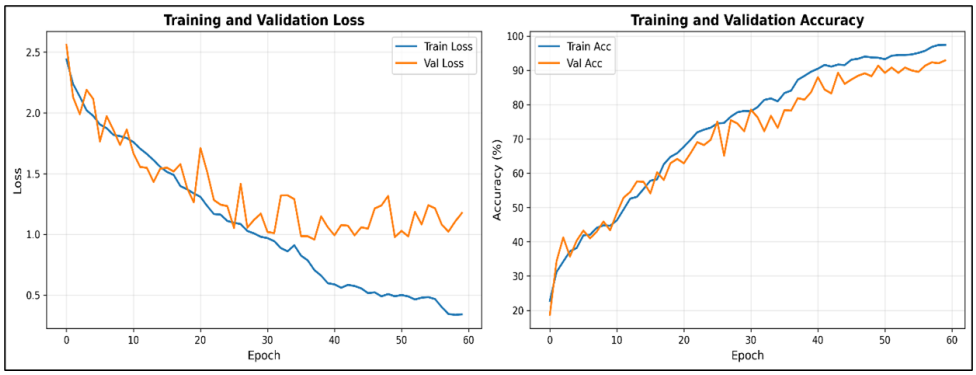

FIGURE 3. Training and Validation Accuracy and Loss Curves over 60 Epochs.

4.2. **Overall Accuracy and F1.** On the test set, our model achieved an overall accuracy of 94.89%. This means out of all emotion clips, 94.89% were correctly classified into the right emotion category. Since accuracy can be misleading in class–imbalanced scenarios, we also report the macro–averaged F1–score, which is the average of F1–scores for each class (treating each class equally regardless of support). The macro F1–score is 94.2%, closely matching accuracy; this high value reflects both high precision and recall across classes as shown in Figure 4.

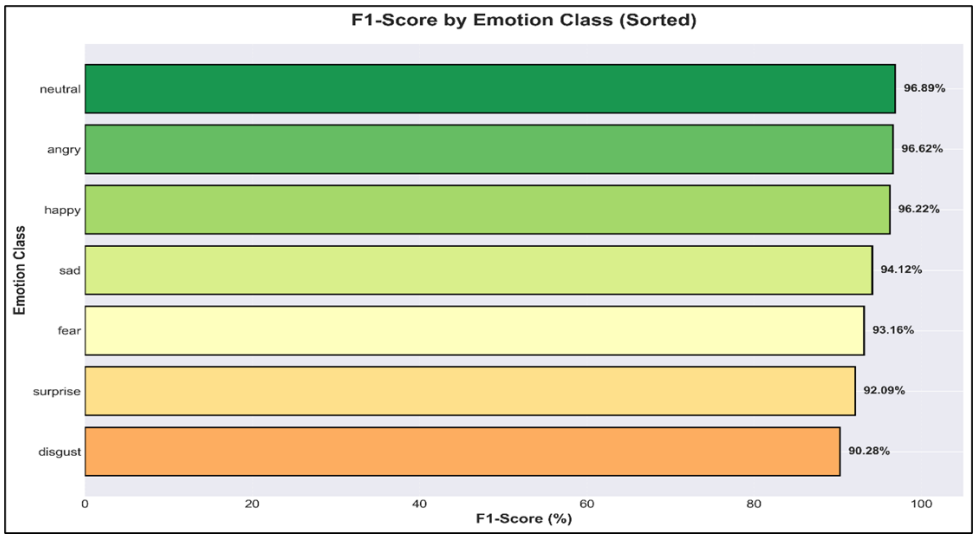

FIGURE 4. Macro-Averaged F1-Score.

These results represent a substantial improvement over previous works on RAVDESS. For context, Jin et al. (2024) report 82.4% accuracy using a bilayer LSTM with attention [4], and the recent open–access study by Salas–Cáceres et al. (2024) achieved 88.11% on RAVDESS using a multimodal LSTM with an Embrace Net fusion [9]. Our model outperforms these points in absolute accuracy. We attribute this gain to our effective attention–fusion mechanism and the regularizing effect of temporal smoothing, as well as careful training and augmentation.

4.3. **Class–wise Performance.** Table 1 and Figure 5 report precision, recall, and F1–score for the seven emotion classes in the RAVDESS test set. The results are well balanced, with all F1–scores staying between 0.90 and 0.96. "Happy" achieves 96.22%, likely because it has clear vocal and facial cues, such as smiles and laughter–like intonation. "Surprise" also performs well at 92.09%, probably because its expressions are more exaggerated and easier to detect. "Fear" reaches 93.16%, although it is sometimes confused with anger or sadness. The merged "neutral/calm" class records 96.89%, with only a small overlap with sadness. Since no class falls below 90% F1, the model appears reliable across all emotions and does not favor one class over another.

TABLE 1. Classification performance on the RAVDESS test set for each emotion category.

Emotion	**Precision (P)**	**Recall (R)**	**F1–score**	**Support (clips)**
Neutral/Calm	95.20	98.64	96.89	221
Happy	97.22	95.24	96.22	147
Sad	96.45	91.89	94.12	148
Angry	95.97	97.28	96.62	147
Fearful	89.38	97.28	93.16	147
Surprised	98.46	86.49	92.09	74
Disgusted	92.86	87.84	90.28	74
Macro Avg.	95.08	93.52	94.20	958
Weighted Avg.	95.00	94.89	94.86	958
Accuracy	94.89	–	–	958

FIGURE 5. Per Class Performance Metrics.

Figure 6 supports this pattern through a radar plot of the F1–scores. The shape is close to a regular heptagon and remains near 1.0 on all axes, which shows that no emotion is clearly weaker than the others. This balance is important for real–world emotion recognition.

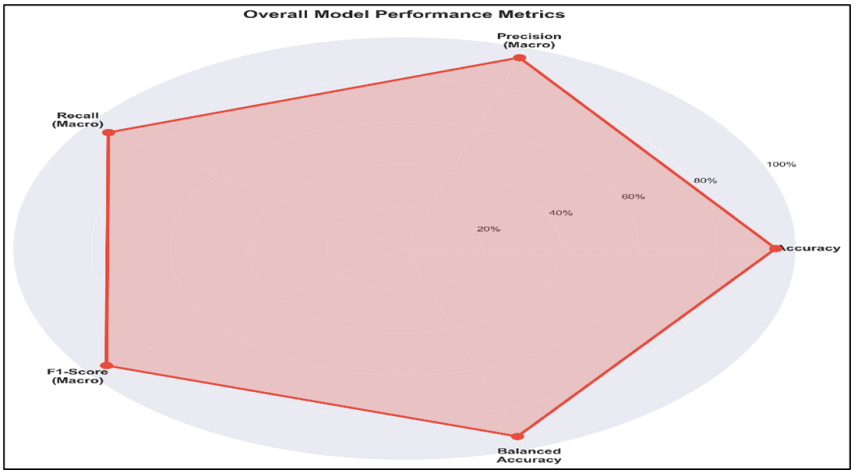

FIGURE 6. Overall Model Performance Metrics in Spider Chart Form.

4.4. **Confusion Matrix.** Figure 7 shows the confusion matrix in raw counts and normalized percentages. Most predictions are concentrated on the diagonal, which means the model classifies most test samples correctly. Recall is especially strong for neutral, angry, and happy. The few errors mainly happen between emotions that are similar in meaning.

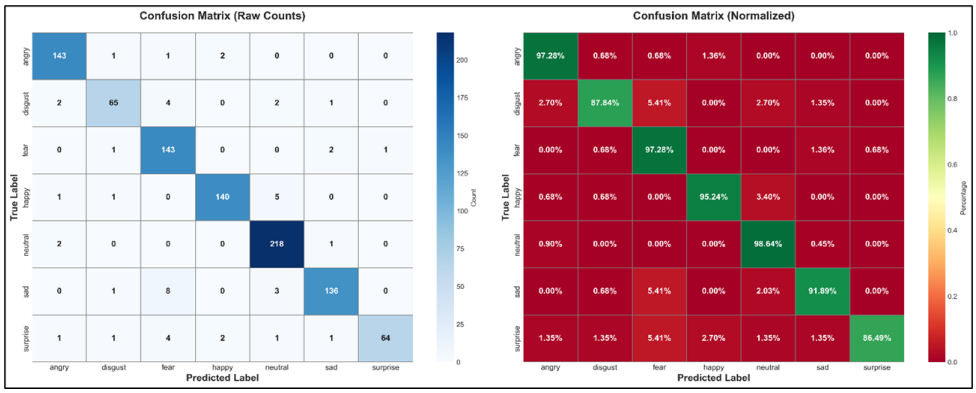

FIGURE 7. Confusion Matrix in Two Forms: Raw Counts (Left) and Row–Normalized Percentages (Right).

4.5. **ROC Curves.** A one–vs–all approach is provided in Figure 8 with ROC curves for each emotion. The curves stay close to the top–left corner, and therefore, the model can detect each emotion correctly with low false alarms. These also maintain very high AUC scores for each class (between 0.988 and 0.998), with an average of about 0.993. To put it more simply, this suggests the classifier separates emotions very well and proves robust while adjusting the decision threshold to either prioritize precision or recall, according to the problem application.

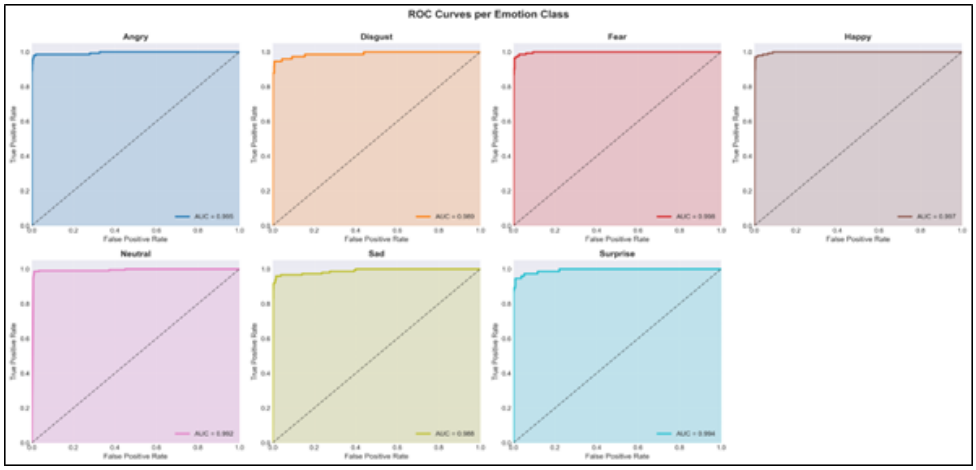

FIGURE 8. ROC curves (one–vs–all) for each emotion class on the test set and Macro–Averaged Area under the ROC Curve (AUC) Across the Seven Emotions.

4.6. **Qualitative Results and Real–World Evaluation.** Noting that not all models performed well over various datasets, this study shows how the model performed in terms of response and learning across models in the real world, and through qualitative examples. Figure 9 shows six samples from the dataset, each featuring a different actor and labeled with the model's predicted emotion. The model appropriately identifies expressions within these individual actors, suggesting the model performs relatively well independently from the actor's identity. This implies that the learned representation is capable of accurately reproducing the important facial attributes of each emotion and generalizing well to many subject areas.

FIGURE 9. Sample model predictions across time for six actors, showing predicted emotion labels aligned with facial frames.

In addition, we test the learned model under real–time application on a child subject, demonstrated in Figure 10. In this demo, the system handles a live video of a child displayed in an unconstrained environment (e.g., a classroom) and provides emotion predictions per frame live. The model tracks a child's changing expressions well, with minimal latency, meaning it stays very stable and responsive in real world applications.

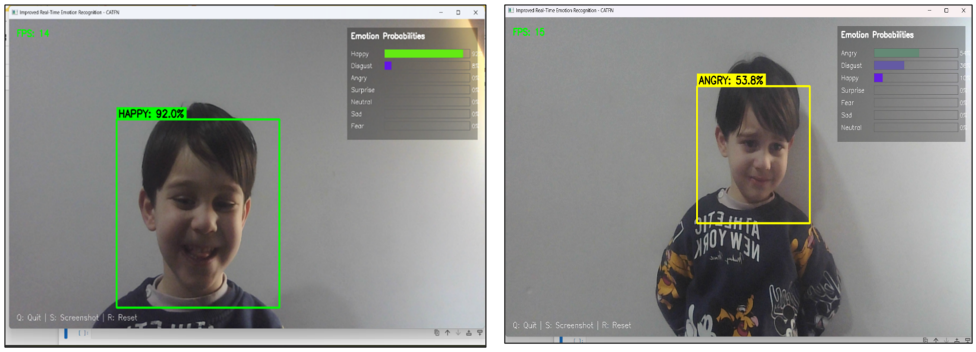

FIGURE 10. Live video stream real–time emotion recognition

5. **Discussion.** Results of our real–time multimodal emotion recognition model are analyzed in this section in terms of (1) class–wise performance patterns, (2) advantages of multimodal fusion and temporal smoothing approach, (3) system limitations, and (4) implications for deployment.

5.1. **Temporal Consistency Smoothing.** This TCS module eliminates short–term fluctuations and enhances prediction stability. In comparison to post–processing filters, our method learns the temporal smoothness during training, yielding more human–like decision timing. While it generates very minimal latency (around 0.2–0.5 s), it significantly enhances reliability and stability for applications such as conversational agents. If adjustment is needed, the smoothing factor will allow faster adaptation.

5.2. **Comparison with Human Performance.** Human recognition on acted datasets such as RAVDESS can be about 90–95%. This is similar to our model, which performs as well as humans and better even with emotions, e.g., fear, where human beings' exaggeration with respect to fear may overshadow subtle cues. Through multi–frame and multi–modal input, it achieves consistent classification where single–cue human judgment can fail.

5.3. **Limitations.** At high accuracy, limitations still lie. RAVDESS includes samples of acted fixed phrases with no natural variation. Since audio synchronization was imperfect from actual demos and was still dependent on auditory input, it became somewhat reliant on visual input. Temporal smoothing could also cause recognition of transient fluctuations of emotions to slip. In addition, attention weights provide little information about the model's internal logic.

5.4. **Modalities Beyond Audio–Visual.** EEG and physiological signals help with emotion inference but require intrusive hardware. We propose a camera–and–mic–based approach with competitive results that is non–intrusive and a lower–key approach that can be competitive in the future, though for our high–precision systems we would expect a more immersive sensory environment.

5.5. **Generalizability and Real–Time Readiness.** Our architecture runs at ∼20 FPS on GPU and is ideal for real–time deployment. Thanks to its modular architecture, these systems can be re–trained on additional datasets. In order to generalize, the system needs to be adapted to create novel emotional experiences that are spontaneous and diverse, as well as optimize audio streaming pipelines.

6. **Conclusion.** This paper introduces a multimodal emotion recognition system that operates in real–time using an attention–guided fusion network of facial and vocal data and an innovative Temporal Consistency Smoothing (TCS) module. Taken together, they address the instability and lack of real–time response in previous affective computing models. This architecture is a fusion of ResNet–based visual features and CNN–LSTM audio encoder to capture complementary input from both modalities. The attention–fusion network learns dynamic weighting among inputs, enhancing robustness over emotion classes. The TCS module combines training loss and inference–time filtering, reduces jitter markedly and generates a ∼0.6% increase in macro F1–score, augmenting the stability and applicability of predictions in real–time. Our model accomplishes 94.89% accuracy and a macro F1–score of 94.2% on seven RAVDESS emotion classes, exceeding previous benchmarks. The model performs well and performs consistently across all classes of emotions, including the rarely distinguishable emotion pairs, such as fear and surprise or sadness and neutrality. The combined methods reveal the advantages of integration. The system also demonstrates impressive promise for immediate use on mobile and embedded platforms, running at about 20 FPS on a GPU. The qualitative results and ROC analysis also confirm that our model is accurate and easy to interpret by this modality–weight attention mechanism. Outside of the technical aspects, the model may cater for other pragmatic applications, such as emotion–aware therapeutic tools in healthcare, adaptive learning systems in education, and systems for interaction with the user as a customer service platform or driver monitoring system. Furthermore, the proposed TCS could likewise be of great value to other, sequential tasks, such as video activity recognition, where temporal consistency is the aim. Taken together, these results mean that real–time emotion recognition can be accurate and stable: it comes closer to natural, coherent responses with thoughtful multimodal design when emotion–aware AI can be taught well.

Acknowledgment. The authors are grateful to the Department of Computer Science under the College of Education for Pure Science at the University of Mosul in Iraq for their support.

REFERENCES

[1] Baltruaitis, T., C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[2] Wani, T. M., et al., "A comprehensive review of speech emotion recognition systems," *IEEE access*, vol. 9, pp. 47795–47814, 2021.

[3] Pan, B., et al., "A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods," *Neurocomputing*, vol. 561, p. 126866, 2023.

[4] Jin, Z. and W. Zai, "Audiovisual emotion recognition based on bi-layer LSTM and multi-head attention mechanism on RAVDESS dataset," *The Journal of Supercomputing*, vol. 81, no. 1, p. 31, 2025.

[5] Dixit, C. and S. M. Satapathy, "Deep CNN with late fusion for real time multimodal emotion recognition," *Expert Systems with Applications*, vol. 240, p. 122579, 2024.

[6] Zhang, S., et al., "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects," *Expert Systems with Applications*, vol. 237, p. 121692, 2024.

[7] Middya, A. I., B. Nag, and S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities," *Knowledge-based systems*, vol. 244, p. 108580, 2022.
[8] Ranganathan, H., S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *2016 IEEE winter conference on applications of computer vision (WACV)*, IEEE, 2016.
[9] Salas-Cáceres, J., et al., "Multimodal emotion recognition based on a fusion of audiovisual information with temporal dynamics," *Multimedia tools and applications*, vol. 84, no. 23, pp. 27327–27343, 2025.
[10] Cambria, E. and B. White, "Jumping NLP curves: A review of natural language processing research," *IEEE Computational intelligence magazine*, vol. 9, no. 2, pp. 48–57, 2014.
[11] Waleed, G. T. and S. H. Shaker, "Speech Emotion Recognition on MELD and RAVDESS Datasets Using CNN," *Information*, vol. 16, no. 7, p. 518, 2025.
[12] Wu, Y., Q. Mi, and T. Gao, "A comprehensive review of multimodal emotion recognition: Techniques, challenges, and future directions," *Biomimetics*, vol. 10, no. 7, p. 418, 2025.
[13] Moorthy, S. and Y.-K. Moon, "Hybrid Multi-Attention Network for Audio–Visual Emotion Recognition Through Multimodal Feature Fusion," *Mathematics*, vol. 13, no. 7, p. 1100, 2025.
[14] Miksik, O., et al., "Efficient temporal consistency for streaming video scene analysis," in *2013 IEEE International Conference on Robotics and Automation*, IEEE, 2013.
[15] Conway, A. M., et al., "Frame–by–frame annotation of video recordings using deep neural networks," *Ecosphere*, vol. 12, no. 3, p. e03384, 2021.
[16] Ramaswamy, M. P. A. and S. Palaniswamy, "Multimodal emotion recognition: A comprehensive review, trends, and challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 6, p. e1563, 2024.
[17] Livingstone, S. R. and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.
[18] Colunga-Rodriguez, A. A., et al., "Developing a Dataset of Audio Features to Classify Emotions in Speech," *Computation*, vol. 13, no. 2, p. 39, 2025.
[19] Tzirakis, P., et al., "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 8, pp. 1301–1309, 2017.
[20] Datta, N., J. Sikder, R. Chakma, and R. K. Das, "Head Features-Based Deep Learning Approach for Recognizing Emotion, Gender and Age," *Journal of Information Hiding and Multimedia Signal Processing*, vol. 14, no. 4, pp. 184–194, 2023.